









Automated Semantic Annotation of Species Names in Handwritten Texts

Lise Stork^{1,2} , Andreas Weber³ , Jaap van den Herik^{1,2} ,
Aske Plaat^{1,2} , Fons Verbeek^{1,2} , and Katherine Wolstencroft^{1,2} 

¹ Leiden Institute of Advanced Computer Science, Leiden, the Netherlands
{l.stork,k.j.wolstencroft,f.j.verbeek,a.plaat}@liacs.leidenuniv.nl

² Leiden Centre of Data Science, Leiden, the Netherlands
h.j.vandenherik@law.leidenuniv.nl

³ University of Twente, Enschede, the Netherlands
a.weber@utwente.nl

Abstract. In this paper, scientific species names from images of *handwritten* species observations are automatically recognised and annotated with semantic concepts, so that they can be used for document retrieval and faceted search. Until now, automated semantic annotation of such named entities was only applied to printed or digital text. We employ a two-step approach. First, word images are classified, identifying elements of scientific species names; **Genus**, **species**, **author**, using (i) visual structural features, (ii) position, and (iii) context. Second, the identified species names are semantically annotated according to the *NHC-Ontology*, an ontology that describes species observations. Internationalised Resource Identifiers (IRIs) are assigned to the elements so that they can be linked and disambiguated at a later stage by individual researchers. For the identification of scientific species names, we achieve an average *F1* score of 0.86. Moreover, we discuss how our method will function in a semi-automated annotation process, with a fruitful dialogue between system and user as the main objective.

Keywords: Deep learning · Ontologies · Taxonomy ·
Scientific names · Semantic annotation · Historical biodiversity research

1 Introduction

Handwritten material brought back from biodiversity expeditions is an important source of information for naturalists and historians. An abundance of these records is available for research [21]. Much of these data, however, remain computationally inaccessible and difficult to explore [4]. This presents an interesting challenge to both the field of *information extraction* and *document retrieval*. Scientific descriptions or depictions of species observations carefully employ the systematic organisation of species variations. Thus, despite the often difficult nature of the data - hard-to-read, multi-lingual, historical texts - document retrieval can exploit the systematic organisation of the document content.

Since the onset of field work in biodiversity expeditions, species observation data have been manually recorded by researchers. Records are fittingly named *field books* [15]. Starting from the first part of the 18th century, Linnaean taxonomy and binomial nomenclature was generally used for the classification and naming of species [18]. Therefore, most historical field books found today in museums and other institutions adhere to Linnaeus's *Systema Naturae* [19]. Due to a common system for the classification of organisms, historical species names can potentially be referenced and compared to current ones, allowing researchers to study the changes in biodiversity over time. However, transforming raw historical biodiversity data to usable structured knowledge is still one of the main challenges of historical taxonomy research [7, 22].

In this work we use state-of-the-art techniques from computer vision and semantic web technologies to (i) identify the elements of scientific species names in handwritten document images, and (ii) link and structure the elements, using an ontology for species observations. We use the MONK handwriting recognition system [23] to segment the document images into single word images. Our main contribution is the identification and semantic annotation of scientific species names from word images containing *handwritten* text. We build on previous work [27], where an ontology and software for semantic annotation of species observation records was constructed and tested with domain experts. Here, we advance these methods by automating the process of semantic annotation. Biological taxonomies, once extracted from field books, can be used by algorithms aiming to exploit query expansion techniques, while it allows users to semantically query, or browse through, field book collections. As the species names are structured via a controlled vocabulary that is well-used in the domain, extracted species names can also be federated across collections.

2 Species Classification and Nomenclature

In the binomial nomenclature, scientific names consist of minimally two and maximally four types of elements. The first type identifies the genus to which the organism belongs. The second type is called the *specific epithet*, the specific species within that genus. Commonly, the binomial is followed by the author name, and the date when the name was published in literature. It is also common for a name to have more than one author. Below in Fig. 1, an example of a scientific species name from a field note is given; it dates back to 1821.

Species names are ambiguous due to evolving taxonomical systems, nomenclature and opposing views within the science of classification [12, 18]. Therefore, scientific names become valuable for scientific research when they are compared to synonyms or homonyms from alternative classifications and their respective meta-data. In the rest of this paper, we will use the term *scientific name* to refer to, minimally, a *genus* and *species* tuple or *genus*, *species* and *author* triple.

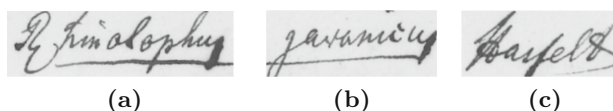


Fig. 1. A scientific name in binomial nomenclature: (a) *Rhinolophus* (genus) (b) *javanicus* (species) (c) *Hasselt* (author of the name: *Johan Coenraad van Hasselt*)

3 Related Work

Organisations and researchers that dedicate themselves to the preservation of natural history collections, such as *IdigBio*¹ or the *Biodiversity Heritage Library* [9], continuously develop new methods to digitise specimen collections in a cost-effective and sustainable way, in order to facilitate ongoing species research. The automatic extraction of scientific names from text is essential for the management of archival resources. Therefore, there are several examples of methods for extracting and disambiguating species names from printed texts, but extracting the same information from handwritten texts is much more of a challenge. Taxongrab [13], for example, automatically extracts species names from printed biological texts. The Biodiversity Heritage Library, that aggregates scans of biodiversity publications and field notes, indexes scientific names extracted from the publications - printed text - in their collection, to improve accessibility for taxonomists. They match the text, extracted via Optical Character Recognition (OCR), with the Taxonomic Name Server (TNS) to identify likely scientific names [9]. They are not the only ones exploiting the power of automatic text processing for the digitisation of natural history collections. Software has been developed to parse OCR output of printed text to formalised Darwin Core² entries for archival and retrieval purposes [10]. Drinkwater and others [8] investigate the aid of OCR in the digitisations of herbarium specimen labels, finding significant increase in time effectiveness using OCR output to sort specimens prior to database submission, and to add data to minimal database records. They explicitly note that OCR is currently only possible for typed and printed labels and not for handwritten text.

Handwritten Text Recognition (HTR) is one of the more challenging tasks within the field of Document Image Analysis and Recognition (DIAR), mainly due to the huge variety in writing styles and languages, paper degradation, overlapping words and historical handwriting. The recognition of named entities - real word objects, such as: *locations*, *persons*, *organisations* - in handwritten text can help document understanding and searchability of the text, and can potentially aid handwriting recognition [5]. Formerly, Named Entity Recognition and Classification (NERC) was a task solely used on digital text [17], but it has just recently also been applied directly to handwritten text [1, 5, 25, 28]. Especially when few instances of words exist and a collection consists of many different

¹ <https://www.idigbio.org/>.

² <http://rs.tdwg.org/dwc/>.

handwritings and connected words, making it difficult to create character-based representations, the identification of key words can help make the text searchable, and potentially aid HTR. Moreover, in many cases, full-text transcriptions of entire pages of field books are not required in order to make them digitally accessible.

In this contribution, we develop a novel approach to identify domain specific named entities, elements of *scientific species names*, in historical *handwritten* document images. Rather than first transcribing the text and performing NERC afterwards on the digital text, we exploit characteristics of the document images to identify the named entities, using terms from the NHC-Ontology³ to classify and organise them. We argue that the ability to quickly index handwritten document images based on scientific names, ranks and authors, helps users to navigate through large collections of documents in online libraries, such as the Biodiversity Heritage Library. It opens up possibilities for faceted search, semantic querying and semantic recommendations. Additionally, maintaining a link to the word image and its position in the full document image is important to allow the repetition of image processing experiments as well as to allow researchers to view the original document and therefore the extracted text in context.

4 Data

One of the main issues history of science and natural history researchers encounter is the inaccessibility of natural history archival collections. Field books, drawings and specimens are physically stored in museum collection facilities or research institutes, hidden from external researchers and policymakers interested in long-term developments of global biodiversity [7].

Table 1. Data set class count

Class	Genus	Species	Author	Other	Total
<i>y</i>	0	1	2	3	
<i>n</i>	177	167	144	17309	17797

Transcribed field books exist online, but (to the best of our knowledge) no segmented and annotated images of handwritten species observations are available online for experimental research using image processing methods. Therefore, word images from 240 field notes from a natural history collection have been segmented and semantically annotated. This has been carried out in the context of the project *Making Sense of Illustrated Handwritten Archives* [31].⁴ From a field book on mammals, we selected field notes from four different writers, to account for different handwriting styles and structures, ensuring a representative

³ <https://makingsense.liacs.nl/rdf/nhc/>.

⁴ <http://www.makingsenseproject.org>.

data set to demonstrate how the automated methods perform on heterogeneous, real-world data. The segmented word images were obtained from a nichesourcing effort, with the help of a handwriting recognition system MONK and a group of domain expert labellers. The word images were subsequently manually annotated using four classes, as shown in Table 1. Two of four classes are taxonomical entities. The third class refers to the publisher of the taxonomical name, and lastly we have the class *Other*, which includes all words that do not belong to any of the previously mentioned classes. The final counts of examples per class are shown in Table 1. The process of labelling and annotating words is time-consuming and, in our case, requires expert knowledge. Therefore, limited training data is available. As machine learning methods generally require a very large number of annotated samples, methods have to be adjusted to the data set size to acquire a predictive model that generalises well. These adjustments are described in Sects. 5 and 6. This is also one of the challenges of such projects; to create an adaptive learning system with a generic method that learns from small amounts of annotated data, but adapts to new data and performs better over time when more data is annotated. The data set used in this work can be found online.⁵

5 Scientific Name Extraction Model

Below we describe the methods that were used in this work. The full pipeline is shown in Fig. 2, the blue rectangle indicating the scope of this work.

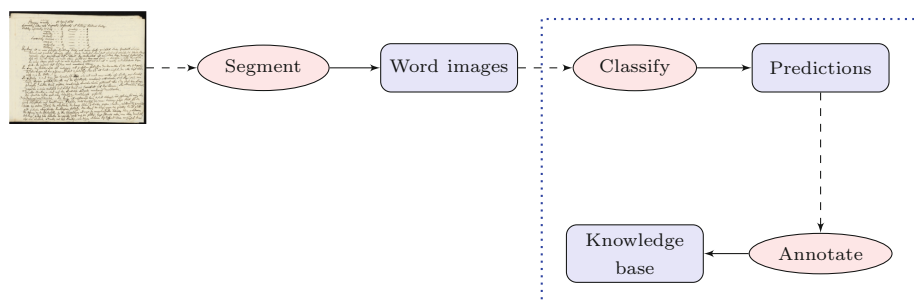


Fig. 2. The full pipeline: automated semantic annotation of scientific names

We used the MONK handwriting recognition system, developed by Schomaker, for word segmentation [3, 23, 29, 30]. First, the system segments handwritten document images into lines and second, relative to those lines, into word zones that potentially hold words. The system allows the labelling, at the word level, of word images by domain experts. It then uses these labels for

⁵ 10.5281/zenodo.2545573.

HTR. In this work, the word images were manually annotated using four semantic concepts, or classes: *genus*, *species*, *author* and *other*. The classification of each word image to its corresponding semantic class is discussed in Sect. 5.1. In Sect. 5.2, we discuss the semantic annotation of the classified word images using the NHC-Ontology⁶ for species observations.

5.1 Classification of Word Images

To classify the word images to one of four classes, we use three distinct features; *visual structural features*, *position* and *context*. We chose to create one single neural architecture, built with help of Keras [16], that could be trained end-to-end, so that the classification error is only propagated once, in contrast to using predictions from multiple classifiers and combining them after training to form a single prediction. The final architecture is explained visually in Fig. 3, and will be discussed below.

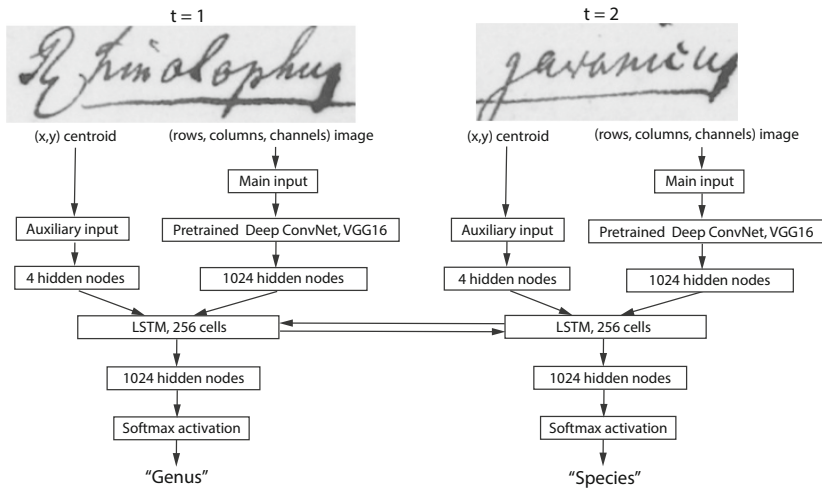


Fig. 3. The MLP-CNN-LSTM architecture, “unrolled” for both time steps t .

Visual Structural Features. The feature detector that was used in this work for the detection of visual structural features is a Convolutional Neural Network (CNN) [14]. It has been shown that CNNs outperform other neural networks on image recognition tasks [26]. The basic network used here is a deep CNN for object recognition developed and trained by Oxford’s Visual Geometry Group (VGG) and called the VGG network [26]. We use their configuration, with 16 convolutional layers, and import weights from the VGG, pre-trained on the ImageNet task [6]. Previous work [20] has demonstrated that transferring image representations with CNNs overcomes the problem of training with limited training

⁶ <http://www.makingsense.liacs.nl/rdf/nhc/>.

data, e.g., less than a few thousand training images, despite differences in image statistics between the *source* data set and *target* data set. By, for instance, training on the ImageNet task, the VGG model learns filters on various different scales, which can be used as feature extractors for other types of images. These features, extracted from handwritten documents with help of the convolutional part of the VGG network, are used for training a simple Multi-Layer Perceptron (MLP) on our task.

Position. In addition to visual features, the position of a word in a document often provides a good descriptive feature for the recognition of a named entity. The position is therefore often used as a feature in the field of NERC, however, it has been used more often in text [17] than in images [1, 5, 28]. In this work, we use the *relative* centroid of a word image’s position in the image as input features to a simple MLP. Hence, each training example $(\mathbf{x}^{(i)}, y^{(i)})$, $\mathbf{x}^{(i)} \in \mathbb{R}^2$, where every x_i lies within the interval $[0,1]$, is used to train a simple MLP with 4 hidden layers. To train the entire model end-to-end, we concatenated the last hidden layers of both models. The merged hidden layer therefore has a size of $1024 + 4 = 1028$.

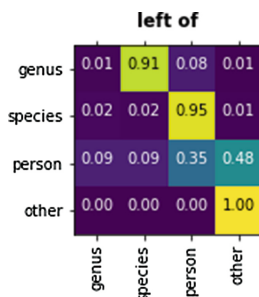


Fig. 4. Adjacency matrix that shows frequencies for word bi-grams (sequences of two adjacent words). E.g., ‘genus’ was **left of** ‘species’ 91% of the time ‘genus’ was encountered.

Context. As a third feature type, we introduce context: the characteristics of adjacent word images, specifically *bi-grams*. Figure 4 shows frequencies for word image bi-grams. First, horizontal pairwise alignment was calculated for each pair of word images $(w^{(i)}, w^{(j)})$, $w \in W$, where $i \neq j$. They were seen as horizontally aligned if $y_1^{(i)} < y_c^{(j)} < y_2^{(i)}$, where $y_1^{(i)}$ indicates the first y coordinate of the bounding box of $w^{(i)}$, $y_2^{(i)}$ the second, and $y_c^{(j)}$ the y coordinate of the centroid of $w^{(j)}$.

Second, the right neighbouring word of $w^{(i)}$ was retrieved by calculating all pairwise vertical distances for the horizontally aligned words: $dist_{ij} = x_c^{(i)} - x_c^{(j)}$, where $x_c^{(i)}$ and $x_c^{(j)}$ refer to the x coordinates of the centroids of $w^{(i)}$ and $w^{(j)}$.

The smallest negative distance indicated right adjacency. The adjacency matrix only takes into account instances that actually *have* an adjacent word, as it could be that a word is surrounded by white space on every side.

As expected, the different classes have strong co-occurrence dependencies. Therefore, we converted the data set to sequences of size two (bi-grams), and added a last layer to the model architecture for sequence prediction. For an adequate prediction we used a Bidirectional Long Short-Term Memory (BLSTM) neural network, a type of Recurrent Neural Network (RNN) that implements a *memory node* in order to learn long-term dependencies between features [24]. By using the bidirectional variant of the LSTM [11], dependencies can be learned in both horizontal orientations, see Fig. 3. This is beneficial for our work, as in the bi-gram *species-author*, the identification of the ‘author’ class largely depends on the visual characteristics of the word image left adjacent to it.

```
nc:taxon1 rdf:type dwc:Taxon
          nhc:scientificNameAuthorship nc:author1
          nhc:taxonRank nc:species

nc:author1 rdf:type foaf:Person

nc:anno1 oa:hasBody nc:taxon1
          oa:hasTarget nc:image1.jpg#xywh=x,y,h,w
          oa:hasTarget nc:image1.jpg#xywh=x,y,h,w

nc:anno2 oa:hasBody nc:author1
          oa:hasTarget nc:image1.jpg#xywh=x,y,h,w
```

Listing 1.1. Example of a semantically annotated species name

5.2 Semantic Annotation of Word Images

The NHC-Ontology⁷ is an ontology for species observations, based on the Darwin Semantic Web (DSW) Ontology, and written in OWL.⁸ The ontology is centered around the description of meta-data relating to the observation of an organism, and allows a researcher to describe to which various taxon groups an organism is identified by a researcher. The model uses the Web Annotation Data Model⁹ to link bounding boxes of word images to their semantic labels. In the exemplary fragment above, listing 1.1, two images refer to a genus and a species, which together constitute one taxonomical name `nc:taxon1`¹⁰ of rank `nc:species`. They are linked to the publisher of the name with the `nhc:scientificNameAuthorship` predicate.

⁷ <http://makiingsense.liacs.nl/rdf/nhc/>, <https://github.com/lisestork/nhc-ontology/>.

⁸ <https://www.w3.org/OWL/>.

⁹ <https://www.w3.org/TR/annotation-model/>.

¹⁰ *nc:* is the prefix for the <http://makiingsense.liacs.nl/rdf/nhc/nc#> namespace.

6 Experiments and Results

To analyse the influence of the three features on the predictive performance of the model, we conducted multiple experiments where we tested the performance of the pre-trained CNN, MLP-CNN and MLP-CNN-BLSTM.

6.1 Experimental Methodology

Before training, the images were scaled by dividing them by 255 so that they would fall within the range $[0-1]$. All images were re-sized to the average image dimensions: $y = 74$, $x = 139$. No data augmentation was used. Based on horizontal adjacency, as explained in Subsect. 5.1, image bi-grams were constructed, sequences of $l = 2$, as input to the BLSTM.

The word images were shuffled, keeping together word images from the same page, and thereafter split into a train and test set. As one word image could occur in two bi-grams, we hereby avoid that word images from the test set were also in the training set, which would bias the classification results. However, by shuffling the pages, we still ensure that the model does not overfit to one writing style or structure. We used 80% of the word images for training and the remaining partition as test set, making sure that 20% of the scientific name elements were in the test set. As classes in the word bi-grams were highly imbalanced, we used random minority oversampling with replacement, to increase the counts of samples from minority classes in the training data. When training a CNN, oversampling is thought to be the best method to deal with imbalanced data sets with few examples in minority classes, and appears to work best if the oversampling totally eliminates the imbalance [2]. However, as we are dealing with sequences rather than singular samples, we chose to oversample sequences, e.g., *species-author*. Converted back to singular images, this would result in a *step imbalance* with a small imbalance ratio $p = \pm 1.1$ rather than a large imbalance ratio of $p = \pm 16$ [2].

The networks were all trained using the Adam classifier with a learning rate of 10^{-4} and categorical cross-entropy loss. Each network was trained using early stopping with patience 2, meaning that training was stopped when, for two epochs, the validation error was increasing. Per epoch, the weights were only stored if the predictive performance had increased compared to the previous epoch. In the testing phase, thresholding was applied to the output of the networks to compensate for oversampling the data during training, as oversampling alters prior probability distributions. One way to perform thresholding is to simply correct for these prior probabilities, by dividing the output of the network for each class, then seen as posterior probabilities, by the estimated prior probabilities. In our case, the imbalance was not completely eliminated, so the thresholds were calculated as the ratio between the original class counts and those after oversampling.

As a final step, the output of the model that performed best was used to test the whole pipeline. Word images from the test set, that were classified as scientific names, were assigned IRIs within the project's namespace, e.g., `nc:taxon1`.

The names were linked and semantically enriched using terms from the ontology and transformed to the Resource Description Framework (RDF) format. The code can be found online.¹¹

6.2 Results and Discussion

Table 2 summarises the final classification results for each network. Due to a large class imbalance, precision and recall were used to assess the predictive power of the classifier. Reporting accuracies would be misleading, as they would portray the underlying distribution rather than the predictive power of the model (if the model would always predict ‘Other’, it would be a bad predictor for the task, but the accuracy would be 93%, as the ‘Other’ class accounts for 93% of the data). The table indicates that the BLSTM produced the highest average *F1* scores for each class. The addition of the BLSTM layer specifically increases precision and recall scores for the author names. This makes sense; without context these appear similar to regular words. The input of centroid data to the network does not have an effect on the recall or precision of author names, but does increase precision for the retrieval of species names. Figure 5 shows 4 images from the test set that were misclassified. While both the CNN and MLP-CNN network misclassify most of the same word images, the output of the MLP-CNN-BLSTM is quite different. Image (a) and (b) were both misclassified by the networks without the BLSTM layer, but were correctly classified by the

Table 2. Classification results per network

Method	Class	Precision	Recall	F1-score	Support
CNN	Genus	0.80	0.78	0.79	36
	Species	0.64	0.97	0.77	33
	Author	0.78	0.78	0.78	32
	Other	1.00	0.97	0.98	525
	avg/total	0.82	0.77	0.80	626
+MLP	Genus	0.85	0.81	0.83	36
	Species	0.81	0.88	0.84	33
	Author	0.78	0.78	0.78	32
	Other	0.99	0.99	0.99	525
	avg/total	0.96	0.96	0.96	626
+BLSTM	Genus	0.86	0.89	0.88	36
	Species	0.94	0.91	0.92	33
	Author	0.78	0.88	0.82	32
	Other	1.00	0.99	0.99	525
	avg/total	0.98	0.97	0.98	626

¹¹ <https://github.com/lisestork/asa-species-names>.

final model. Image (a) for example, was classified as ‘Species’, while actually being labelled as an author name. Visually, it resembles a species name; it is underlined and appears in a similar position on the page. Without context of other words it is challenging to correctly classify such images without proper historical knowledge of the domain. Image (b) was misclassified as ‘Other’, but correctly identified as an author name in the BLSTM model, most likely due to the visual characteristics of the word image that is left adjacent. On the other hand, image (c) and (d) are together misclassified as a species name and its author by the BLSTM network, while they were correctly classified by the other networks. Eyeballing the images, we see that they are adjacent and visually resemble these classes (capitals, underlining).

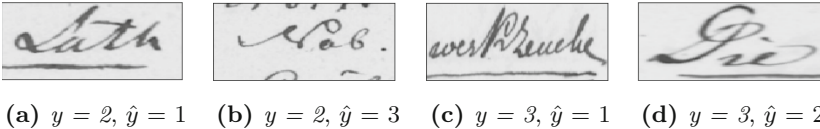


Fig. 5. Four misclassified examples. Classlabels relate to those discussed in Table 1

In Table 3, we present retrieval scores for the identification of complete scientific names from field book pages. A python script parsed the recognised species elements from the test set, and connected them together using the NHC-Ontology. A total of 27 out of 36 species names were retrieved, with an *F1* score of 0.86. Interestingly, there were no false-positives among the final predictions. Figure 6 shows one of the correctly classified scientific names. The final RDF data set can be queried through our online SPARQL endpoint.¹²

Table 3. Final classification results for the detection of scientific names

Method	Class	Precision	Recall	F1-score	Support	Total
+BLSTM	Scientific names	1.0	0.75	0.86	27	36

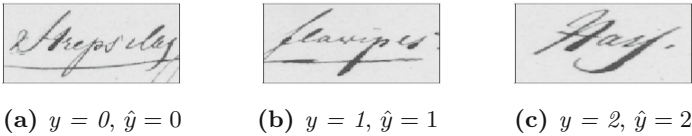


Fig. 6. A correctly classified scientific species name: (a) *Genus* (b) *Species* (c) *Person*

¹² <http://makingsense.liacs.nl/rdf4j-server/repositories/SN>, can be queried through a query editor such as: <https://yasgui.org/>.

6.3 A Semi-automated Process

This work serves as a step within the development of an adaptive system, with the MONK handwriting recognition system at its core [31], for the segmentation, recognition and semantic annotation of handwritten words, named entities and illustrations from historical biodiversity collections. Using labelling input from domain experts, representations of the document images are learned in order to generate new, machine learned, labels. Simultaneously, domain experts can provide contextual knowledge on specific biodiversity expeditions from which the annotation process can benefit. For example, named entities - such as author names - can be used to pre-populate the knowledge base so that they can be retrieved during the semantic annotation process. Moreover, domain experts can link the - validated - automatically identified scientific names to word images containing higher ranks, so that collections can be browsed using faceted search.

7 Conclusions and Future Work

In this work we show that we can accurately identify and classify components of handwritten species observation records from different features: visual structural features, position and context. We show that our methods are applicable even though the data set contains four authors with different handwriting styles and different processes of recording their species observations. A major challenge of working with handwritten text is its irregularity. Our results show that we can mitigate this challenge by building up multiple pieces of evidence for classification by learning from multiple features. Each of the different features we examine in our model adds information and improves the overall results. In addition, as the results are extracted and structured in RDF as part of the process, they are immediately available for search and comparison with other archives - historical or present day.

The entire data set used for these experiments is part of the same expedition archive. Although we represent multiple authors and styles, the next step would be to demonstrate the generic nature of our results by analysing biodiversity records from other expeditions. Once we establish that, we will extend our methods to identify other common classes from biodiversity data, for example, locations, dates and anatomical entities. In the context of the Making Sense project, we aim to integrate the new methods with established methods for automated handwriting recognition, using the MONK system.

Acknowledgements. This work is supported by the Netherlands Organisation for Scientific Research (NWO), grant 652.001.001, and Brill publishers.

References

1. Adak, C., Chaudhuri, B.B., Blumenstein, M.: Named entity recognition from unstructured handwritten document images. In: 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 375–380. IEEE (2016)
2. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **106**, 249–259 (2018)
3. Bulacu, M., van Koert, R., Schomaker, L., van der Zant, T.: Layout analysis of handwritten historical documents for searching the archive of the cabinet of the Dutch queen. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition, vol. 1, 2, pp. 357–361. IEEE (2007)
4. Canfield, M.R.: *Field Notes on Science & Nature*. Harvard University Press, Cambridge (2011)
5. Carbonell, M., Villegas, M., Fornés, A., Lladós, J.: Joint recognition of handwritten text and named entities with a neural end-to-end model. *arXiv preprint arXiv:1803.06252* (2018)
6. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Proceedings of the 2009 Computer Vision and Pattern Recognition, CVPR, pp. 248–255. IEEE (2009)
7. Drew, J.A., Moreau, C.S., Stiassny, M.L.: Digitization of museum collections holds the potential to enhance researcher diversity. *Nature Ecol. Evol.* **1**(12), 1789–1790 (2017)
8. Drinkwater, R.E., Cubey, R.W., Haston, E.M.: The use of optical character recognition (OCR) in the digitisation of herbarium specimen labels. *PhytoKeys* **38**, 15–30 (2014)
9. Gwinn, N.E., Rinaldo, C.: The biodiversity heritage library: sharing biodiversity literature with the world. *IFLA J.* **35**(1), 25–34 (2009)
10. Heidorn, P.B., Wei, Q.: Automatic metadata extraction from museum specimen labels. In: Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications, pp. 57–68 (2008)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
12. Kennedy, J.B., Kukla, R., Paterson, T.: Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In: Ludäscher, B., Raschid, L. (eds.) *DILS 2005*. LNCS, vol. 3615, pp. 80–95. Springer, Heidelberg (2005). https://doi.org/10.1007/11530084_8
13. Koning, D., Sarkar, I.N., Moritz, T.: Taxongrab: extracting taxonomic names from text. *Biodivers. Inf.* **2**, 79–82 (2005)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
15. MacGregor, A. (ed.): *Naturalists in the Field*. Brill, Leiden (2018)
16. Chollet, F., et al.: *Keras* (2015). <https://keras.io>
17. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the Seventh Conference on Natural Language Learning, vol. 4, pp. 188–191. Association for Computational Linguistics (2003)
18. Miracle, M.E.G.: On whose authority? Temminck’s debates on zoological classification and nomenclature: 1820–1850. *J. Hist. Biol.* **44**(3), 445–481 (2011)
19. Müller-Wille, S.: Names and numbers: “data” in classical natural history, 1758–1859. *Osiris* **32**(1), 109–128 (2017)

20. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717–1724 (2014)
21. Page, L.M., MacFadden, B.J., Fortes, J.A., Soltis, P.S., Riccardi, G.: Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience* **65**(9), 841–842 (2015)
22. Sarkar, I.N.: Biodiversity informatics: organizing and linking information across the spectrum of life. *Briefings Bioinform.* **8**(5), 347–357 (2007)
23. Schomaker, L.: Design considerations for a large-scale image-based text search engine in historical manuscript collections. *It - Inf. Technol.* **58**(2), 80–88 (2016)
24. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**(11), 2673–2681 (1997)
25. Shi, Z.: Datefinder: detecting date regions on handwritten document images based on positional expectancy. Master's thesis, University of Groningen, Groningen, the Netherlands (2016)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556 (2014)
27. Stork, L., et al.: Semantic annotation of natural history collections. *Web Semant. Sci. Serv. Agents World Wide Web* (2018). <https://doi.org/10.1016/j.websem.2018.06.002>
28. Toledo, J.I., Sudholt, S., Fornés, A., Cucurull, J., Fink, G.A., Lladós, J.: Handwritten word image categorization with convolutional neural networks and spatial pyramid pooling. In: Robles-Kelly, A., Loog, M., Biggio, B., Escolano, F., Wilson, R. (eds.) *S+SSPR 2016. LNCS*, vol. 10029, pp. 543–552. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49055-7_48
29. Van der Zant, T., Schomaker, L., Haak, K.: Handwritten-word spotting using biologically inspired features. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(11), 1945–1957 (2008)
30. van Oosten, J.-P., Schomaker, L.: Separability versus prototypicality in handwritten word-image retrieval. *Pattern Recogn.* **47**(3), 1031–1038 (2014)
31. Weber, A., Ameryan, M., Wolstencroft, K., Stork, L., Heerlien, M., Schomaker, L.: Towards a digital infrastructure for illustrated handwritten archives. In: Ioannides, M. (ed.) *Digital Cultural Heritage. LNCS*, vol. 10605, pp. 155–166. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75826-8_13