Artificial intelligence vs human handwriting: annotating damaged manuscripts

Dumitru Scutelnic[1], Laura Gazzani[2], Paolo Pellegrini[3], Claudia Daffara[4]
[1]Dept. of Computer Sciences, University of Verona, Italy - dumitru.scutelnic@univr.it
[2] Dept. of Computer Sciences, University of Verona, Italy - laura.gazzani@univr.it
[3] Dept. of Cultures and Civilizations, University of Verona, Italy - paolo.pellegrini@univr.it
[4]Dept. of Computer Sciences, University of Verona, Italy - claudia.daffara@univr.it

ABSTRACT (ENGLISH)
To decipher the text of damaged manuscripts using state-of-the-art AI methods such as Deep Learning, a training phase with an annotated dataset is required. We have developed an optimized dataset to assess the potential of AI in different book heritage applications, ranging from text recovery to diagnostic analysis, using high-resolution multispectral imaging. This dataset is continuously expanded with image stacks acquired on ongoing case studies; here, as exemplary cases, a damaged notebook by a 20th-century Italian writer and a manuscript by a 19th-century Italian intellectual. The annotation process must yield precise and detailed labels for the text and for each individual handwritten character, thereby providing AI models with suitable input for training. Unlike standard annotation approaches, which rely primarily on transcribing the text, we propose a different method based on instance segmentation of each character in the manuscript. Masks are traced to follow the exact shape of each character, each assigned to its own distinct class. Specific annotation criteria were defined in cross-disciplinary collaboration with philologists, physicists, and AI experts to maximize the system's potential for accurate handwriting recognition in degraded materials.
Keywords: multispectral imaging; damaged manuscripts; characters annotation; semantic segmentation, deep learning.

ABSTRACT (ITALIANO)
Intelligenza artificiale vs scrittura umana: annotare manoscritti danneggiati.
Per decifrare il testo di manoscritti danneggiati usando lo stato dell'arte dei metodi AI, come il Deep Learning, è necessaria una fase di training su dataset annotato. Abbiamo sviluppato un dataset ottimale per testare il potenziale dell'AI in varie applicazioni che riguardano il patrimonio librario, dal recupero del testo alla diagnostica, usando l'imaging multispettrale ad alta risoluzione. Il dataset viene continuamente ampliato con serie di immagini acquisite su casi studio; qui, come casi esemplari, un taccuino danneggiato di uno scrittore italiano del XX sec. e un manoscritto di un intellettuale italiano del XIX sec. Per fornire all'AI un modello adeguato, il processo di annotazione deve produrre "label" precise e dettagliate del testo e di ogni carattere della scrittura a mano. A differenza del normale metodo di annotazione, basato sulla trascrizione del testo, proponiamo un approccio diverso basato sulla segmentazione di istanze per ciascun carattere del manoscritto. Le maschere vengono tracciate seguendo ogni carattere, ognuno dei quali ha una sua classe distinta. I criteri specifici di annotazione sono stati determinati in collaborazione interdisciplinare, tra filologi, fisici, ed esperti di AI, in modo da massimizzare il potenziale per il riconoscimento della scrittura manoscritta anche in materiali degradati.
Parole chiave: imaging multispettrale; manoscritti danneggiati; annotazione dei caratteri; segmentazione semantica e di istanza, deep learning.

1. INTRODUCTION
Text recognition using artificial intelligence (AI) has evolved rapidly in recent years, thanks to the advancement of Deep Learning (DL) methods (Wang et al., 2021). The methods initially were based on network architectures such as Convolutional Neural Networks (CNNs) for recognizing single characters or words for different handwriting and alphabets (Liu et al., 2011). The introduction of Transformers (e.g., TrOCR and HTR-VT) (Minghao Li et al.,2021; Yuting Li, 2024), transformed traditional optical character recognition (OCR) into sophisticated systems capable of handling complex text formats such as a whole line of text. The solutions analyzed offer excellent performance when the text is clear and legible. However, when manuscript materials are degraded and are characterized by partially or completely illegible areas, such techniques show significant difficulties in character recognition. This has been verified

on our datasets using established OCR tools, such as Transkribus, eScritporium, and OCR4All. Undamaged manuscripts (such as Del Bene's case study) show better results when the text is still legible, the ink is uniform, on a uniform and clear background, and the contrast is high. Traditional OCR algorithms fail to provide a transcript when the ink is faded, and reflectance variations alter the contrast against an uneven background due to damaged materials. Consequently, there is a need for a robust strategy to separate ink from noise before text recognition. Similar studies have focused mainly on Asian languages (Maheswari, Maheswari & Aakaash, 2024; Shikha, Sonu & Vivek, 2020), but not as many have examined highly degraded manuscripts, the Latin alphabet or the Italian language. It is worth mentioning the Vesuvius Challenge concerning the decipherment of Greek text in the carbonized papyrus scrolls known as the Herculaneum papyri (Lourenco et al. 2023).

In this regard, the AIPAD Project (Artificial Intelligence and Physics for Art Diagnostics) focuses on the analysis and diagnostics of damaged historical manuscripts with non-invasive techniques and advanced data processing. Among the tasks are: 1) to exploit multispectral imaging in the UV-VIS-IR range to detect manuscript materials; 2) to annotate the multispectral dataset and train AI for handwritten text recognition on degraded manuscripts.

The dual nature of the manuscript – intangible text and physical support – is the key point. Optical imaging allows the different layers to be inspected, providing additional information (Perino et al., 2024; Tonazzini et al., 2019). It is known that UV imaging can reveal surface features of the manuscript (faded ink, erased text), while IR imaging can detect subsurface features (underwriting, carbon-based inks on blackened parchment). Recently, an integrated protocol with multiband imaging up to the mid-IR range has proven effective in the study of damaged manuscripts (Cimino et al., 2022).

In this study, we employ a multispectral imaging dataset acquired in the UV-VIS-NIR (up to 1 micron), spatially registered (Mazzocato et al., 2024). The dataset, called MADAM (Carcagnì et al., 2024) consists of 90 patches, each one a multispectral stack of 19 images. MADAM is continually populated with new images acquired on ongoing case studies; here, we consider a multispectral stack of a flood-damaged notebook of a 20th-century Italian writer, where the writing is degraded and not clearly visible to the naked eye. Multispectral imaging maximized the information in the 660 nm band, where the ink is more detectable, as shown in Fig. 1. An additional example includes a set of VIS images from a manuscript of Benedetto Del Bene (1749-1825), family annals, which are clearly legible (Fig. 2). They were included to enable predictive models to effectively learn and generalize character recognition. The challenges for handwritten character recognition are the diversity of writing styles, inconsistent spacing between letters and lines, and uncertainty about the number of lines and characters in a patch. Thus, we evaluated different strategies of annotation from simple transcription to more complex ones with masks that include from character level to the even non-uniform background.

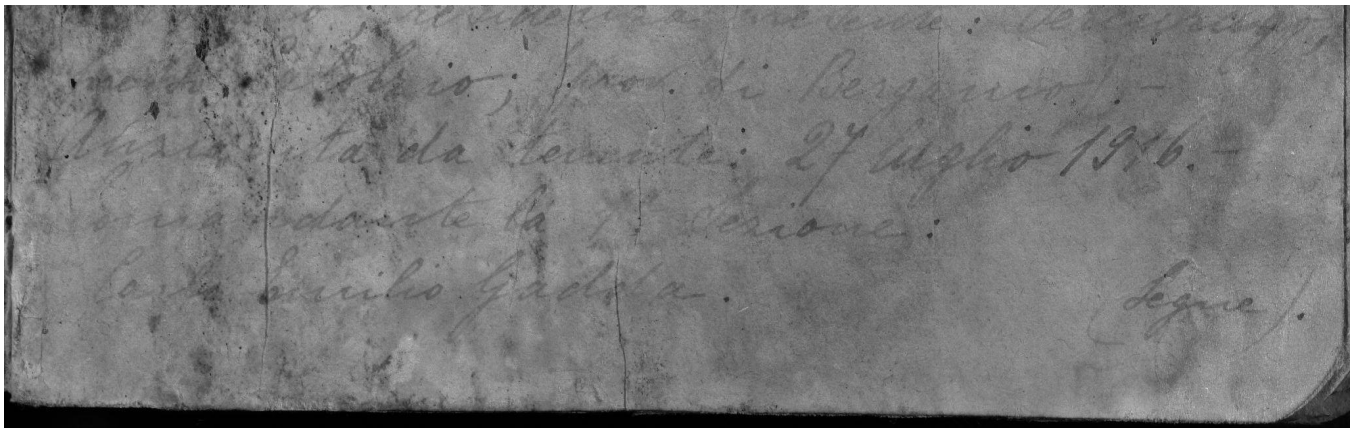

Figure 1: Image at 660 nm of an exemplary patch from the highly degraded manuscript.

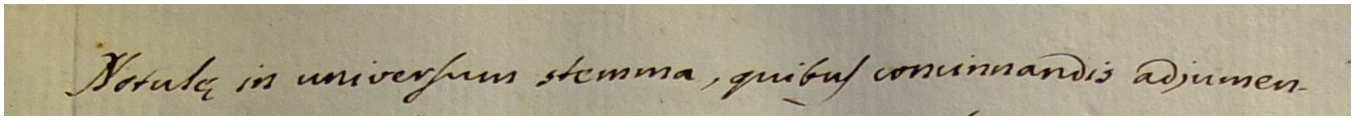

Figure 2: VIS image of an exemplary patch of Benedetto Del Bene's manuscript, page 1.

2. CRITERIA FOR ANNOTATION

Annotation must be as detailed and precise as possible to achieve accurate and scalable recognition results. Image binarization techniques are adopted to overcome problems of color gradients and different levels of ink transparency. In case of degraded materials, the key point is to consider only the traits of characters and exclude the background, i.e., the spurious signs of the damage, which cause OCR to fail. As annotation technique, we adopted the instance segmentation of each character, by creating masks that provide pixel by pixel the shape of the object, its position, and its class. This method replicates the exact shape of each character on the label by tracing it manually with a drawing pad. The definition of the classes to be annotated was done in synergy with philologists. It was decided to dedicate a separate class to every single character that can be found on the manuscript: one class for every letter (majuscule and minuscule), one for every number and one for every sign of punctuation (period, comma, hyphen, parentheses) – with an average of 70 classes per manuscript (Maheswari et al., 2024). There may be small variations from one manuscript to another, because annotation is language-specific (Maheswari et al., 2024). For example, the Del Bene manuscript (used for testing, written in Latin) presents the character "æ", which is absent in the other dataset. Any illegible characters or any characters that do not have a corresponding key are annotated as the "_" label. Such gaps will be filled by AI in the transcription, using a statistical approach based on the semantic content of the text, and possibly by the hypotheses of philologists, who would act as the ultimate quality check for the annotations (to fine-tune the algorithm) and the AI predictions and transcriptions. At the request of the philologists who are lending their expertise to this project, the annotation is carried out with particular attention to the ligatures between the letters to make the characters recognizable in any possible context. This is aimed at facilitating a successful recognition and transcription for an AI algorithm. This is also done to support any studies regarding the script itself, since it would be easier to pick and compare all the different ways of writing the same letter. Some examples of annotated patches can be found in Fig. 3.

(a)

(b)

Figure 3: VIS image of Del Bene manuscript (a) and image at 660 nm (b) of patches annotated by using the method of semantic segmentation with multiple classes (multiple colours of the characters).

This annotation procedure is very laborious and requires a lot of time and resources, however, the advantages are in scalability and possibility to use new DL techniques with more complex architectures. The annotation includes also bounding boxes that can be exploited for object detection, not fully explored in the context of text recognition. One aspect to consider concerns bounding boxes that may include parts of the background with signs of degradation that can confuse the DL model. Overcoming this problem would require huge datasets and resources for the AI training.

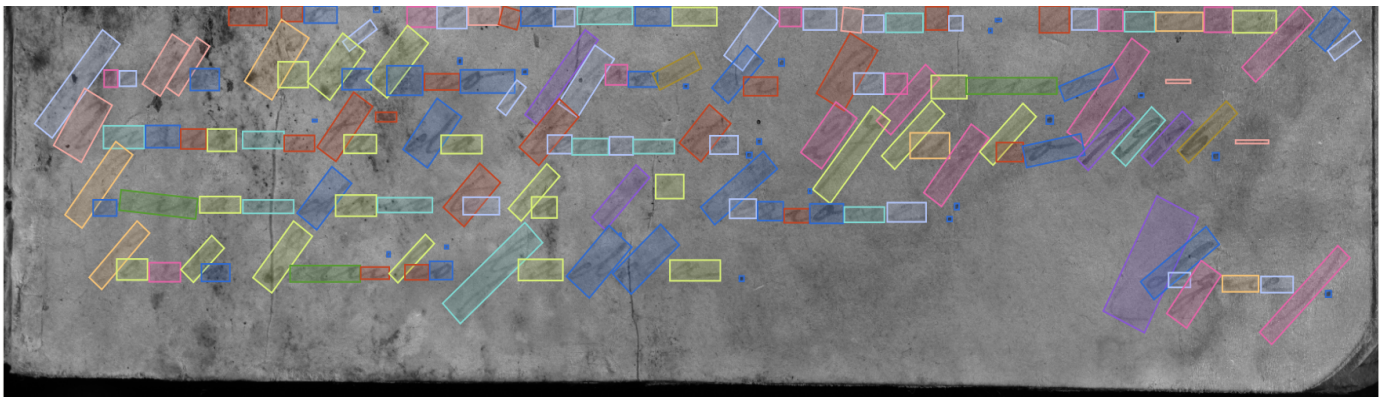

Figure 3: Image at 660 nm of the patch annotated by using the method of object detection with bounding boxes.

3. ARCHITECTURE AND METRICS

As DL models for instance segmentation, various architectures such as Mask R-CNN, U-Net, PointRend, HTC and DeepLabv3+, etc. (Chen et al., 2017; Singh et al., 2024), never used in text recognition, will be evaluated. While, for object detection, architectures such as Faster-RCNN, YOLO, Single Shot MultiBox Detector (SSD), RetinaNet, etc. (Srivastava et al., 2021), which are little explored for text recognition, will be evaluated. In addition, the procedure will be integrated with robust and established OCR methods such as Convolutional Recurrent Neural Network (CRNN) (Puarungroj 2020; Singh et al., 2024) with CTC Loss and Transformer OCR (e.g., TrOCR and HTR-VT) (Minghao Li et al.,2021; Yuting Li, 2024) to further improve text recognition.

Quality control is ensured by the expert philologists who accurately annotate each letter and background. While, the mean Average Precision (mAP), Intersection over Union (IoU), Recall, Precision and F1-score were used as metrics to estimate the learning performance of the models. For text recognition, Character Error Rate (CER), Word Error Rate (WER), Word Accuracy (correct words / total words) and CTC Loss (Yuting Li, 2024; Retsinas et al. 2024) will be used.

4. (EXPECTED) RESULTS

Fig. 4 represents the groundtruth obtained through the annotation of masks for each character. This is the expected result using the innovative approach of instance segmentation, where, in addition to the recognition of the text traced by the segmentation, its location is also identified together with its class. The approach aims to improve character recognition capability over available OCR tools, which do not perform well in the case of damaged manuscripts.

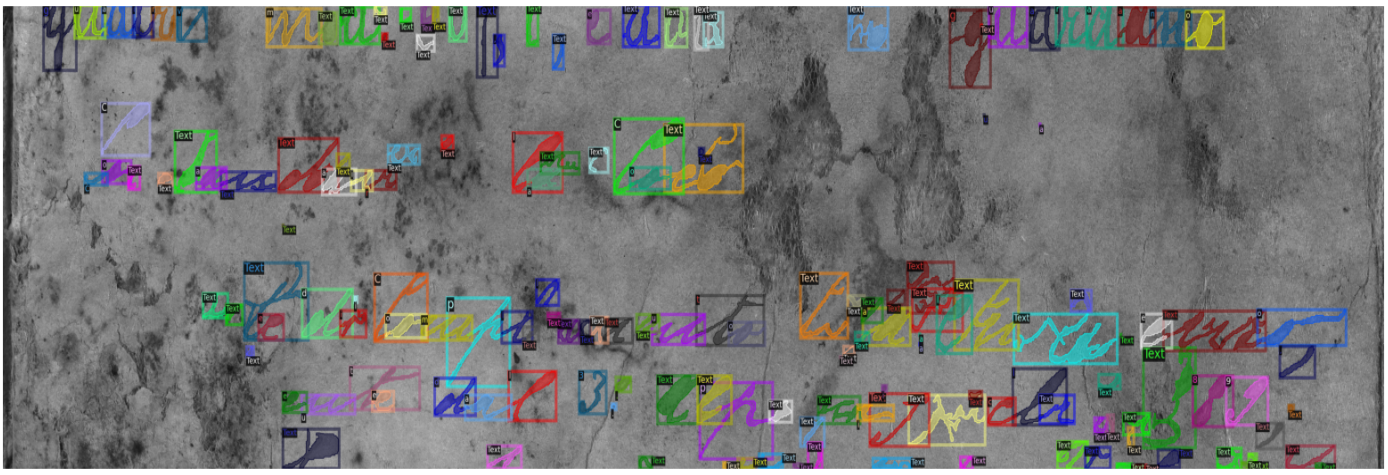

Figure 4: Annotations superimposed on a patch taken from the same page; the corresponding letters are indicated on any annotation, in the corner of the polygons.

5. FUTURE WORK

Future developments of the project include the expansion of the multispectral dataset with further historical damaged manuscripts from libraries in Verona. Another interesting direction for further research is the annotation of the background for diagnostic purposes by mapping the manuscript degradation features. Another task, which is underway, is the analysis and annotation of the physical aspects of the manuscripts that are intentional, such as watermarks and ruling. More expansions can also be directed towards new alphabets, such as the Cyrillic and the Greek alphabet.
In addition, the dataset will be made accessible to the community with impact in the field of text-assisted restoration using advanced optical and artificial intelligence techniques and in the development of new predictive methods.

ACKNOWLEDGEMENTS

This research is funded by the European Union – NextGenerationEU, M4C2 component, investment 1.1, PRIN PNRR 2022 Project: "Artificial Intelligence and Physics for Art Diagnostics (AIPAD): the killer application of the historical manuscripts and palimpsests", P2022SFEPN - CUP B53D23029240001 - Grant Assignment Decree No. 1372 (1/9/2023) by the Italian MUR.

REFERENCES

Yintong Wang, Wenjie Xiao and Shuo Li, Offline Handwritten Text Recognition Using Deep Learning: A Review, ICAACE 2021, Journal of Physics: Conference Series, 1848 (2021) 012015 https://doi.org/10.1088/1742-6596/1848/1/012015.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio2, Cha Zhang, Zhoujun Li, Furu Wei, TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models, Computation and Language, 2021, https://doi.org/10.48550/arXiv.2109.10282

Yuting Li, Dexiong Chenb, Tinglong Tanga, Xi Shenc, HTR-VT: Handwritten Text Recognition with Vision Transformer, Computer Vision and Pattern Recognition, 2024, https://doi.org/10.48550/arXiv.2409.08573

Maheswari, S.U., Maheswari, P.U. & Aakaash, G.R.S. (2024). An intelligent character segmentation system coupled with deep learning based recognition for the digitization of ancient Tamil palm leaf manuscritps. *Heritage Science 12*, 342. https://doi.org/10.1186/s40494-024-01438-4.

Shikha, C., Sonu, M., Vivek, S. (2020). Ancient text character recognition using deep learning. *International Journal of Engineering Research Technology, 13(9)*, 2177-2184. https://doi.org/10.37624/IJERT/13.9.2020.2177-2184.

Alex Lourenco, Brent Seales, Christy Chapman, Daniel Havir, Ian Janicki, JP Posma, Nat Friedman, Ryan Holbrook, Seth P., Stephen Parsons, and Will Cukierski. Vesuvius Challenge - Ink Detection. https://kaggle.com/competitions/vesuvius-challenge-ink-detection, 2023. Kaggle.

Michela Perino, Lucilla Pronti, Candida Moffa, Michela Rosellini and Anna Candida Felici, New Frontiers in the Digital Restoration of Hidden Texts in Manuscripts: A Review of the Technical Approaches, Heritage 2024, 7, 683–696. https://doi.org/10.3390/heritage7020034

Liu, C., Yin, F., Wang, D., & Wang, Q. (2011). CASIA Online and Offline Chinese Handwriting Databases. *International Conference on Document Analysis and Recognition*, 37-41. https://nlpr.ia.ac.cn/2011papers/gjhy/gh39.pdf.

Tonazzini, A., Salerno, E., Abdel-Salam, Z.A., Harith, M.A., Marras, L., Botto, A., Campanella, B., Legnaioli, S., Pagnotta, S., Poggialini, F. & Vincenzo Palleschi (2019). Analytical and mathematical methods for revealing hidden details in ancient manuscripts and paintings: A review. *Journal of Advanced Research*, 17, 31-42. https://doi.org/10.1016/j.jare.2019.01.003.

Cimino, D., Marchioro, G., De Paolis, P., Daffara, C. (2023). Evaluating the integration of Thermal Quasi-Reflectography in manuscript imaging diagnostic protocols to improve non-invasive materials investigation, *Journal of Cultural Heritage 62*, 72-77. https://doi.org/10.1016/j.culher.2023.04.009.

S. Mazzocato, D. Cimino, C. Daffara, Integrated microprofilometry and multispectral imaging for full-field analysis of ancient manuscripts, Journal of Cultural Heritage, Volume 66, 2024, Pages 110-116, ISSN 1296-2074, https://doi.org/10.1016/j.culher.2023.11.014.

Carcagnì, P., Del Coco, M., Leo, M., Gazzani, L., Malagodi, M., Paturzo, M., Daffara, C. (2024). MADAM: Manuscript Annotated Dataset based on Multispectral Imaging for Handwritten Text Enhancement and Restoration. Proc. of the International Conference Florence HeriTech, 2024, Firenze. (in press)

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A.L. (2017). Deep Lab: Semantic Image Segmentations with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. https://arxiv.org/pdf/1606.00915.

Puarungroj, W., Boonsirisumpun, N., Kulna, P., Soontarawirat, T., Puarungroj, N. (2020). Using Deep Learning to Recognize Handwritten Thai Noi Characters in Ancient Palm Leaf Manuscripts. Ishita, E., Pang, N.L.S., Zhou, L. (Eds.) *Digital Libraries at Times of Massive Societal Transition. ICADL 2020. Lecture Notes in Computer Science*, Vol 12504. https://doi.org/10.1007/978-3-030-64452-9_20.

Srivastava, S., Divekar, A.V., Anilkumar, C. *et al.* Comparative analysis of deep learning image detection algorithms. *J Big Data* **8**, 66 (2021) https://doi.org/10.1186/s40537-021-00434-w

Sukhdeep Singh, Sudhir Rohilla, Anuj Sharma, An inclusive review on deep learning techniques and their scope in handwriting recognition, Computer Vision and Pattern Recognition, 2024, https://doi.org/10.48550/arXiv.2404.08011

George Retsinas, Giorgos Sfikas, Basilis Gatos, Christophoros Nikou, Best Practices for a Handwritten Text Recognition System, arXiv, 2024, https://arxiv.org/pdf/2404.11339v1