Does Context Matter ? Enhancing Handwritten Text Recognition with Metadata in Historical Manuscripts*

Benjamin Kiessling[1,*,†], Thibault Clérice[2,†]

[1]*École Pratique des Hautes Études, Université PSL, 4-14 rue Ferrus, 75014, France*
[2]*Inria, 48 rue Barrault, 75013 Paris*

Abstract

The digitization of historical manuscripts has significantly advanced in recent decades, yet many documents remain as images without machine-readable text. Handwritten Text Recognition (HTR) has emerged as a crucial tool for converting these images into text, facilitating large-scale analysis of historical collections. In 2024, the CATMuS Medieval dataset was released, featuring extensive diachronic coverage and a variety of languages and script types. Previous research indicated that model performance degraded on the best manuscripts over time as more data was incorporated, likely due to over-generalization. This paper investigates the impact of incorporating contextual metadata in training HTR models using the CATMuS Medieval dataset to mitigate this effect. Our experiments compare the performance of various model architectures, focusing on Conformer models with and without contextual inputs, as well as Conformer models trained with auxiliary classification tasks. Results indicate that Conformer models utilizing semantic contextual tokens (Century, Script, Language) outperform baseline models, particularly on challenging manuscripts. The study underscores the importance of metadata in enhancing model accuracy and robustness across diverse historical texts.

Keywords

handwritten text recognition; medieval manuscripts; metadata

1. Introduction

The digitization wave of the past two decades has significantly increased online access to historical manuscripts. Despite this progress, a substantial number of these documents are available only as images, lacking machine-readable text. Handwritten Text Recognition (HTR) has emerged as a vital tool for converting these images into text, facilitating the analysis of vast historical collections such as Camps et al.'s work [1]. Consequently, multiple large datasets have emerged in recent years [2, 3, 4, 5]. However, most of these datasets are mono- or bilingual, with relatively limited geographical, temporal, scribal, and generic diversity. While this does not affect the quality of the datasets per se, it limits the generalization of models derived from them. Specifically, such models may face vocabulary limitations in the case of language or generic unicity (e.g., corpora composed solely of biblical content [6]), and graphical interpretation issues due to the lack of scribal, temporal, or geographical variation.

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark
*Corresponding author.
†These authors contributed equally.
benjamin.kiessling@ephe.psl.eu (B. Kiessling); thibault.clerice@inria.fr (T. Clérice)
0000-0001-9543-7827 (B. Kiessling); 0000-0003-1852-9204 (T. Clérice)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The Middle Ages, spanning approximately ten centuries, encompass a period of immense linguistic and cultural diversity. This era witnessed the evolution of numerous languages and dialects, each with distinct characteristics and scripts. From Old English and Latin to Old High German and Old French, the linguistic landscape of the medieval period was dynamic and continually evolving. This diversity poses both challenges and opportunities for HTR, as models must be capable of handling a wide array of scripts and languages that changed significantly over time. Addressing these challenges requires datasets that reflect the rich and varied nature of medieval manuscripts, incorporating a broad spectrum of geographical, temporal, and scribal variations to enhance the robustness and generalizability of HTR models.

In late 2023 and early 2024, the publication of the CATMuS Kraken model [7] and subsequently the CATMuS Medieval dataset [8] has opened up new opportunities for training and evaluating generic models across a vast diversity of categories and traits. With 200 manuscripts in their initial release in January 2024, and 250 in their 1.5.0 July release, encompassing 10 languages and 6 other metadata fields, these resources provide a robust framework for developing generalizable models that account for these specific features. However, in their initial study, Pinche et al. [7] indicate that the new generalizing models, trained on the comprehensive dataset, exhibited a drop in performance compared to earlier, more language-specific models. This finding seems to contradict the intended benefit of large, intercompatible datasets[1].

One promising approach to mitigate these issues is the enrichment of handwritten text datasets with metadata. Metadata provides contextual information that can enhance model training and improve recognition accuracy. For instance, metadata on the century of production, language, script, and genre can help models better understand and adapt to the specific characteristics of the text they are processing.

This paper explores the potential need for metadata-enriched handwritten text datasets. We hypothesize that incorporating detailed metadata can improve HTR performance, particularly for complex historical texts. By analyzing the performance of current models on metadata-enriched versus non-enriched datasets, we aim to demonstrate the benefits of this approach and propose a framework for its implementation.

2. Background and Related Work

Automatic text recognition in general and in particular the processing of historical typewritten and machine-printed material has seen a stellar rise in recent years. This advancement has had a profound impact on scholarly work, especially in the field of historical research. The retrodigitization and accurate transcription of most types of historical documents, which were once laborious tasks, can now be accomplished with relative ease and sufficient precision to enable a multitude of novel investigations.

Metadata and domain knowledge have long played important roles in the design of automatic text recognition systems (ATR). In fact, the limitations of early ATR methods, principally utilized for the processing of documents in tightly constrained domains, necessitated incorporating

[1]It is important to note, however, that the models were compared using a similar architecture, without any hyperparameter optimization based on the newly acquired diversity of the dataset. This suggests that further optimization and adaptation may be necessary to fully leverage the potential of such diverse datasets.

both to restrict the search space and boost accuracy to acceptable levels. Examples of these are systems designed to aid in automatic letter sorting where the vocabulary is effectively closed but also general-purpose ATR software such as Tesseract [9] utilizing extensive dictionaries and other means of language modelling.

Unfortunately traditional techniques to incorporate metadata have strong normalizing tendencies which are problematic for the recognition of historical documents which often have diverse language use, orthography, and multilingualism. While modern ATR software with its more powerful text recognition methods dispenses with many of these accuracy-boosting techniques, this is doubly true for software designed for historical document digitization like [10] which in most cases go to great lengths to eliminate them as far as possible.

Automatic Text Recognition The principal paradigms employed in typical Automatic Text Recognition text recognizers have been stable for more than a decade although considerable research has resulted in recognition methods that are significantly more powerful, with higher accuracy, better generalization, and increased ease-of-training than the basic algorithm proposed in [11]. These recognizers are placed at the end of a pipeline of interconnected processes. A rudimentary but fairly standard ATR pipeline will ingest a digital scan of a page image at a time, perform any necessary pre-processing, e.g. rectification, dewarping, or binarization, find individual lines on the page image in a step called layout analysis, and feed the identified lines individually through the text recognizer. In a final step, the recognition results of the individual lines are reassembled into a paginated text by concatenation and serialization into raw text files or combined with data from the layout analysis to produce a digital facsimile, most frequently in standardized formats like ALTO or PageXML.

The most important feature of these ATR systems is that they implement text recognition as a sequence to sequence modelling task where the input sequence is typically a line image and the desired output sequence a string of characters. There are multiple ways to construct such a sequence-mapping text recognizer albeit the most popular way is with Connectionist Temporal Classification loss (CTC) [11] which permits the model to learn without requiring an explicit alignment between input and output. Further, these methods have multiple other advantages, some especially pertinent for historical document retrodigitization: training data creation is typically much faster than with older character-based ATR methods as line-wise annotation is generally more efficient, a lack of explicit character segmentation markedly improves error rates on cursive writing and connected scripts, and the ability of the recognizer to take contextual information into account boosts accuracy of characters that are difficult to recognize in isolation, e.g. in the case of degraded writing.

Style-aware HTR and other metadata-enriched architectures While interventions contributing domain knowledge into ATR systems at a general language level, e.g. with dictionaries or language modelling, are widespread, approaches explicitly leveraging other metadata that might be known about the text to be recognized have rarely been described in the literature. Minor exceptions include a method described in [12], similar to the semantic context token in section 3.2, for the processing of standardized European Accident Statements, achieving a 10% reduction in CER with an architecture concatenating a metadata vector to the encoder features

in a standard CNN-LSTM trained with CTC.

[13] describes a metadata-aware handwritten text recognition method albeit for a very different use case. A k-shot learning algorithm for style-aware HTR based on meta-learning, a base model is first trained from a text recognition training set enriched with writer labels where each meta-learning task corresponds to writing produced by a single writer. During inference on writing produced by a previously unknown individual scribe, an update of the model weights with a low number of labelled samples results in an adapted model for this particular scribe. This approach boost accuracy by around 5-7 percentage points in comparison to naive fine-tuning.

Automatic Text Recognition (ATR) datasets for historical, and specifically medieval, manuscripts likely began with Latin script datasets from the Historical Databases of IAM, notably the Partzival [14] and St. Gall [15] subsets. These datasets, which remain widely used for benchmarking new ATR engines, are relatively small (1,000 and 4,000 lines respectively), derive from single source documents, and are fundamentally incompatible due to differing annotation guidelines.

Late 2010s datasets, such as those developed by D. Stutzmann and the company Teklia [16, 5, 4][2], have taken a more focused generic approach (e.g., cartularies, books of psalms) and provided a significantly larger number of lines (more than 120,000 for HIMANIS). However, these datasets are limited by their generic and language unicity, and their use of annotation guidelines that resolve abbreviations restricts their reusability in multilingual settings. This is due to genre- or language-specific abbreviations and normalizations, which pose challenges for contextual-dependent abbreviation resolution [2].

The CATMuS dataset offers an innovative framework to address these limitations, enabling testing of ATR models across diachronic (8th-16th century), diageneric (from practical documents to poetry), and multilingual (10 languages) variations. With a consistent annotation approach, the CATMuS dataset [8] allows for the development and evaluation of single models capable of handling the rich diversity of medieval manuscripts.

3. Proposed method

We propose two basic approaches to evaluate the impact of metadata on recognition performance at different points of a text recognition method and evaluate it against a baseline of an advanced attentional text recognizer based on the Conformer architecture [17] and the default hybrid convolutional and recurrent neural recognizer of the kraken OCR engine. Although our experiments are run on an adaptation of fairly complex Conformer models the fundamental idea can be employed in almost any type of text recognizer based on neural networks.

3.1. Text Recognition with Transformers

The baseline system consists of an adapted Conformer, a Transformer-style [18] neural network augmented with convolutional layers, currently the dominant neural network architecture in

[2]Their publication date is relatively older than their original availability.

automatic speech recognition (ASR). While ASR and ATR share many of the same features, e.g. relatively low-dimensional inputs and a prevalent sequence-to-sequence paradigm, there is no reported use of them in the ATR domain as of yet.

While the fundamental architecture requires no adaptation for text recognition, the size of even very large text recognition datasets is significantly smaller than the corpora of spoken speech typically used in ASR research which necessitates downscaling the network for reliable convergence (encoder_dim $= 144$, encoder_layers $= 16$, num_attention_heads $= 4$). In addition, we adopt the computationally more efficient depthwise-convolution downsampling schema (conv_channels $= 32$, subsampling_factor $= 4$) from [19] which roughly doubles inference speed without accuracy losses.

Our baseline recognizer consists of this down-sized Conformer encoder followed by a single fully connected layer as a decoder. Like most text recognition methods it is trained with CTC loss.

3.2. Semantic Context Token

Our first proposed method explicitly supplies the text recognizer with contextual information of the line to be recognized during training and inference. Given an input image of a line $I \in \mathbb{R}^{w \times h \times c}$ with height h, width w, c channels to the recognizer and a vector $\vec{t} \in \{0, 1\}$ containing the encoded metadata, which we will call the semantic context token, we simply expand the token to size $w \times |\vec{t}|$ and concatenate it to the input resulting in a new input to the network $I' \in \mathbb{R}^{w+|\vec{t}| \times h \times c}$. The neural network is then trained as usual with CTC loss.

The chosen metadata is encoded into semantic context token $\vec{t}$ through a simple multi-hot encoding, suitable for a wide-range of tag-type metadata, placing a high value at a particular position in the vector to indicate the presence of a tag. Classes are dealt with through expansion, e.g. for a language metadata field and possible values $L = \{\text{Castilian}, \text{Venetian}, \text{Latin}\}$ we would be converted into a semantic context token $|\vec{t_L}| = 3$.

An obvious drawback of this method is that the text recognizer needs to be supplied the same array of metadata during both training and inference, i.e. it can only effectively recognize unknown text lines when the same metadata using during training is known.

3.3. Auxiliary Loss

In contrast to the first approach which is intended to induce the recognition model to context switch based on explicitly provided information during inference, our second method relies on an auxiliary loss during training to aid the network in learning the structure of the input data without requiring a semantic context token during inference.

Instead, the network is trained to reconstruct the semantic context token as the output of a side-branch of the text recognition network. This side branch, situated just after the Conformer encoder, consists of a simple adaptive max pooling and fully connected layer and operates on the totality of the encoder features. For a context token t and a prediction of the side branch $\hat{t}$ of size n the auxiliary loss $\mathcal{L}_{\text{aux}}$ is computed using binary cross-entropy (BCE):

$$\mathcal{L}_{\text{aux}}(t, \hat{t}) = -\sum \{l_1, \ldots, l_n\}^\top \tag{1}$$

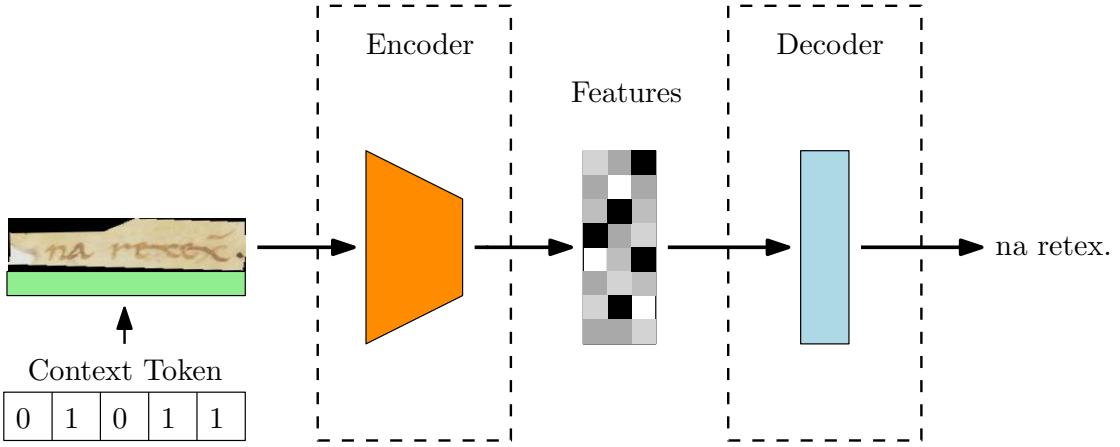


Figure 1: Architecture of the semantic context token method: the multi-hot encoded token is concatenated (light green) column-wise to the input image. The combined input is then fed through the recognition network as normal, both during training and inference. The encoder (orange) is our modified Conformer network, the decoder (light blue) is a single layer feed-forward network.

where

$$l_n = -\left[\hat{t}_n \cdot \log t_n + (1 - \hat{t}_n) \cdot \log(1 - t_n)\right] \tag{2}$$

The overall training objective thus becomes:

$$\mathcal{L} = (1 - w) \cdot \mathcal{L}_{\text{CTC}} + w \cdot \mathcal{L}_{\text{aux}} \tag{3}$$

where w is an additional hyperparameter of the training process that determines the proportion between the main CTC and auxiliary BCE loss. In line with common practice and confirmed with preliminary experiments we chose to put a relatively low weight ($w \in (0.1, 0.3)$) on the auxiliary loss during training.

4. Data

For the purpose of this paper, we utilized the CATMuS Medieval dataset, adhering to the provided dataset splits, which segment the training, validation, and evaluation sets by document. The training and validation splits were sourced from the 1.0.1 release, while the evaluation split was taken from the 1.5.0 release[3] for testing purposes (see Table 8). This approach allowed us to benefit from the expanded and more varied test set, enhancing the robustness of our evaluation without compromising the integrity of our initial training and validation processes.

Representing the diversity, or lack thereof, in the CATMuS dataset is challenging due to the various metrics (lines, characters, pages, or documents) and numerous features to consider (genre, language, script, century, etc.). Language can be seen as a super-category, which is then

[3]We leveraged the release of a larger, more diverse test set for evaluation; however, due to the short time frame (less than five days) between the release of version 1.5.0 and the submission deadline of this paper, we were unable to retrain and redo all experiments. While some documents seem to have undergone metadata correction in between releases, we expect it to have a relatively small impact on our evaluation scores.

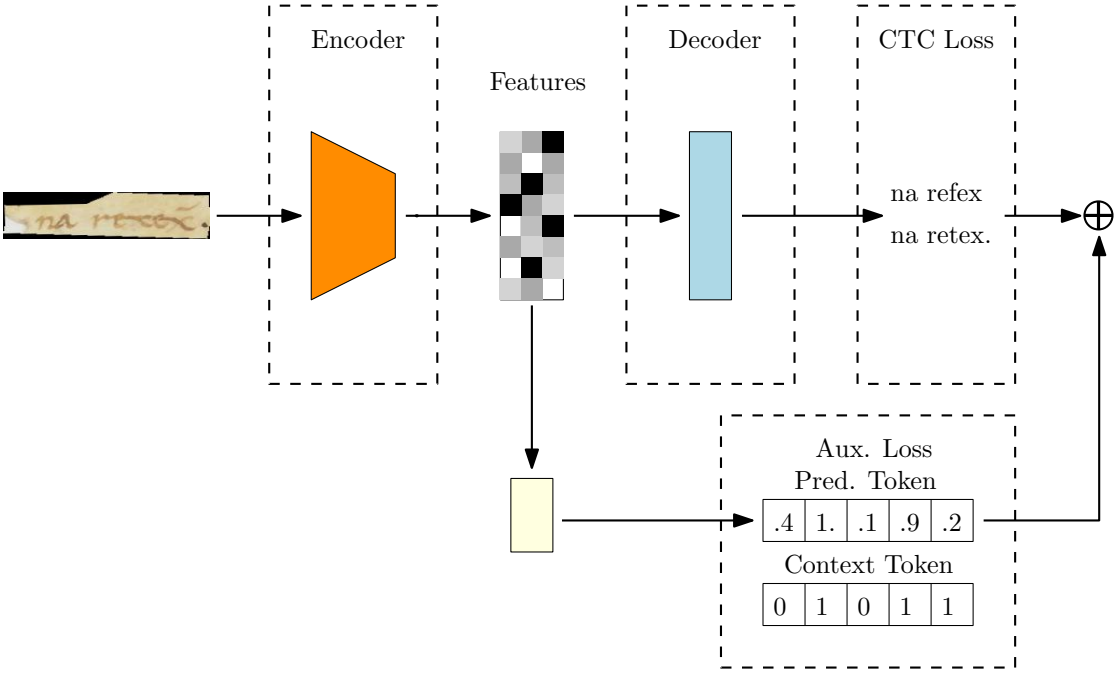


Figure 2: Architecture of the auxiliary loss method: during training the encoder features are processed by the side branch (light yellow) to predict the context token for a particular line. The auxiliary loss $\mathcal{L}_{\text{aux}}$ is merged with the main CTC loss $\mathcal{L}_{\text{CTC}}$ computed on the predicted text to arrive at the overall loss $\mathcal{L}$.

refined by genre if we view genre as primarily limiting vocabulary. In our dataset description, we focus on script (which can serve as a proxy for century), language, and use lines as the metric of choice. Lines are ultimately the unit used for training (sample and batch size) and offer a compromise between document and character count. However, it is important to note that some documents are heavily represented in terms of lines, while others have much longer lines (particularly in the context of prose vs. poetry), affecting the overall representation.

CATMuS 1.0.1 and 1.5.0 are heavily uneven across categories. In Table 2, we identify four particularly challenging "couples" in the test set: 156 lines of Castilian in Humanistica script, 273 lines of French in Semihybrida, 736 lines of Navarese, and 147 lines of Venetian in Textualis script. Each of these scripts has representatives in the training and development sets in other languages, but Venetian has only two documents in CATMuS (1 in train and 1 in test since CATMuS 1.0.0) and Navarese has only one document overall, and only in the test set. However, the Textualis script, which represents these languages, is the most common script in the training and development sets (see Table 1). We anticipate these test lines to be the most difficult for the model to predict. Latin is the most represented language across scripts, missing representation in only five classes in the test sets. Additionally, two scripts (Personal and Print) and two languages (Catalan and English) are absent from the test set entirely. Caroline and Praegothica scripts are overly represented in the test set in terms of lines, but this metric hides a reality for Caroline in number of documents, as three documents in Latin Caroline are in the test set, but

Table 1
Number of lines in train and development split in CATMuS 1.0.0

	Castilian	Catalan	English	French	Italian	Latin	Middle Dutch	Venetian
Caroline			538			6706		
Cursiva	300	482		7560	595	1394		
Gothic						525		
Documentary Script								
Humanistica					929	598		94
Hybrida	7089	196		271	184	1619		
Personal						151		
Praegothica						816		
Print	5552			11308		1880		
Semihybrida	613	172				605		
Semitextualis	9669	416		416		679		
Textualis	7609			28688	444	5922	45998	

Table 2
Representation of lines by couple script-language in the test set in comparison to the data in train and development splits, as a percentage, such that $v = |Lines_{\text{test}}|/(max(1, |Lines_{\text{train}}| + |Lines_{\text{dev}}|))$. When there are no data in the train and development sets, the percentage is normalised using 1 as the number of lines, and values are put in bold.

	Castilian	French	Italian	Latin	Middle Dutch	Navarrese	Venetian
Caroline				101.2			
Cursiva		18.1		60.3			
Gothic Doc.				20.2			
Humanistica	**15600.0**		54.0	45.7			
Hybrida	7.6						
Praegothica				106.1			
Print		3.7					
Semihybrida	25.1	**27300.0**					
Semitextualis				28.4			
Textualis		5.2		4.8	2.3	**73600.0**	**14700.0**

22 different small documents represent this script in the train and dev split[4].

5. Experiments

We perform experiments on the latest 2024 version of the CATMuS Medieval dataset. While this dataset is sufficient in size to train a Conformer model from scratch, the models in our experiments were fine-tuned from a base model trained on around 2.5 million text lines in a

[4]This is another example of how difficult it is to represent the diversity and over-representation of some categories.

large number of scripts and languages in order to reduce the time and computational resources expended.

5.1. Implementation Details

All experiments are performed using the same hyperparameters and identical initial seeds. The model architecture follows section 3.1. Line images are scaled to a fixed height of 96 pixels and padded on both sides with 16 pixels.

The batch size is set to 32, the maximum supported by our Nvidia A40 GP under BFloat16 mixed precision.

Models are trained using the AdamW optimizer [20] for 100 epochs with a cosine learning rate schedule with linear warmup over 35000 iterations, equivalent to slightly more than 8 epochs on our dataset and batch size. Initial learning rate after warmup is $3e-4$ decaying to $3e-5$ by the end of the schedule. The network is regularized with weight decay ($1e-5$), dropout (0.1), and augmentation with random blurring, scaling, rotation, and elastic transforms[5]

5.2. Experimental Setup

Table 3: Selected metadata fields and values

Field	Values
Language	Italian, English, French, Castilian, Latin, Middle Dutch, Navarrese, Venetian, Catalan
Script type	Caroline, Cursiva, Gothic Documentary Script, Humanistica, Hybrida, Praegothica, Personal, Print, Semihybrida, Semitextualis, Textualis
Century	8, 9, 10, 11, 12, 13, 14, 15, 16

We chose to evaluate our methods on a subset, shown in Table 3, of the line-level metadata provided by the CATMuS dataset. To determine the impact of each metadata field and potential synergistic effects on recognition accuracy, both methods were trained with language, script type, and age fields both separately and jointly. For the auxiliary loss weight an upper limit was determined empirically, from below which the values $\{0.1, 0.2, 0.3\}$ were sampled for evaluation.

All models are evaluated on character accuracy. For comparison, baseline models were trained with both the default configuration of the Kraken OCR engine (CNN+LSTM recognizer) and the unmodified Conformer architecture.

6. Results

General results. Out of the two proposed architectures, only the Conformer model using contextual input tokens with all context tokens (Century, Script, Language) consistently outperforms the other models. Specifically, this model surpasses the baseline Conformer architecture, which itself outperforms the original Kraken baselines (see Table 4). Models that utilized a single category of features, such as Language or Century, ultimately performed worse than the baseline. The auxiliary loss approach yielded unexpected results: out of the 12 configurations (four types of tasks with three types of loss weights), half did not converge and resulted in

[5]The source code for all experiments can be found under a *libre* Apache 2.0 at https://github.com/mittagessen/conformer_ocr.git.

Table 4
Test Results (Character Accuracy): Models or combinations not reported failed to converge, exhibiting a micro-accuracy below 15%. Macro-accuracy represents the mean of document-level accuracies.

	Input	Micro-Accuracy	Macro-Accuracy
Kraken		86.56	88.75
Conformer		90.32	92.07
Aux. Loss	All (0.1)	89.74	91.61
	All (0.2)	89.70	91.60
	All (0.3)	89.59	91.31
	Language (0.3)	89.89	91.58
	Script (0.1)	89.96	91.70
	Script (0.3)	89.57	91.43
Context. Input	All	**91.14**	**92.86**
	Century	87.53	89.59
	Language	88.37	90.21
	Script	87.73	89.91

character accuracies below 15%. Even worse, the observed unstable training behavior seems to be unrelated to the chosen weight w, which indicates that optimal hyperparameters must be determined for each new dataset and metadata token.

Accuracy dispersion across manuscript. The Contextual Input model consistently outperforms all other models, with the lowest median CER and the lowest variance. For the most challenging manuscript, it achieves over a 2 percentage point increase in accuracy compared to the Conformer baseline (see Table 5). Additionally, the Contextual Input model, without ablation, exhibits the smallest variance among all models (see Figure 3a). Compared to the baseline (see Figure 3b), the model utilizing the contextual token demonstrates superior accuracy, with a median improvement of 0.64 percentage points. It only underperforms on three manuscripts: Paris, BnF, fr. 6447 (baseline: 97.20%, -0.33); Paris, BnF, lat. 17903 (baseline: 80.25%, -0.32); and Paris, BnF, lat. 130 (baseline: 97.31%, -0.08).

Table 6: Test zoom-in on manuscripts with an unknown language (character accuracy).

	Input	BnF, esp. 65	BnF, ita. 783
Kraken		91.94	90.37
Conformer		93.60	92.91
Context. Input	All	94.08	**93.11**
	All (zeroed)	**94.14**	93.07

Ablation study. To evaluate the impact of the contextual token, we present results with null contextual tokens in Table 7. For models utilizing a single category of contextual input, removing the contextual token results in decreased accuracy, with macro-accuracy dropping by up to 3.2 percentage points for the model using the Century metadata and by as little as 0.88 points for the model using scripts. These findings suggest

Table 5
Test results for the two worst performing manuscripts across models (*Bibiothèque Inter-universitaire de la Sorbonne, 193* & *BnF, Lat. 17903*), and the two best one (BnF, fr. 13496 & BnF, fr. 574). Only the best Aux. Loss and Context. Input are kept.

	Input	BIUS 193	BnF, Lat. 17903	BnF, fr. 13496	BnF, fr. 574
Kraken		77.19	77.27	95.88	94.94
Conformer		80.52	80.25	96.53	97.72
Aux. Loss	All (0.1)	78.64	80.62	**97.43**	97.35
Context. Input	All	**82.22**	79.93	97.25	98.04
	All (zeroed)	82.21	**80.63**	97.08	**98.08**
	Century	74.64	77.69	95.79	95.68
	Language	75.60	78.76	96.30	96.98
	Script	71.32	78.46	95.86	96.78

that the models may be overfitting to the contextual token, as evidenced by the baseline Conformer models outperforming them.

However, for the model using all contextual inputs (Context Input All), the removal of the context token leads to a smaller reduction in efficiency. Despite being less efficient with null contextual tokens, the model still leverages learned features during decoding, aligning with our expectations for the Auxiliary Loss training architecture. The minimal variation in accuracy between the zeroed-out context and the full context (< 0.15 percentage points) while still surpassing the baseline may indicate that the model has effectively learned to separate features, even without manually provided context.

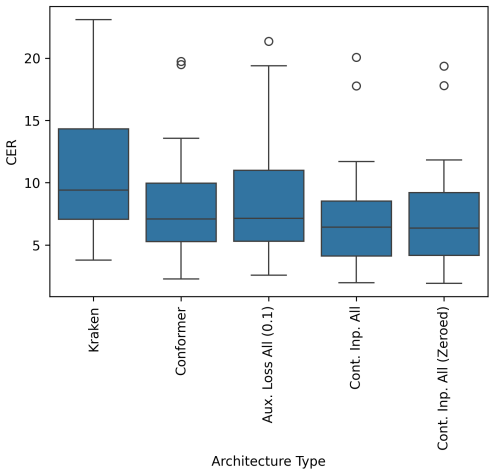
Table 7
Ablation results (character accuracy): All Conformer models using contextual inputs include configurations with the nullification of the contextual token, indicated as (zeroed).

	Input	Micro-Accuracy	Macro-Accuracy
Conformer		90.32	92.07
Context. Input	All	**91.14**	**92.86**
	All (zeroed)	91.13	92.79
	Century	87.53	89.59
	Century (zeroed)	85.52	88.64
	Language	88.37	90.21
	Language (zeroed)	86.55	88.66
	Script	87.73	89.91
	Script (zeroed)	87.22	89.03

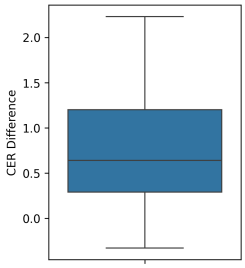
Impact of unknown features. In documents featuring unknown or extremely rare features, such as the Navarrese language (unknown) and Venetian language (represented by only one training sample), our results not only remain stable but also surpass those of the conformer

model when utilizing all contextual tokens. Particularly noteworthy are manuscripts BnF 65 and BnF ita. 783 (*cf.* Table 6), where we observe consistently stronger performance. Even in cases with null semantic tokens, we achieve improvements ranging from +0.2 to +0.4 points in accuracy.

7. Conclusion



(a) Dispersion of the CER across manuscripts per model for the main models



(b) CER difference between the baseline and the best model (Context. Input All non-zeroed).

Figure 3: Dispersion of CER across models on the test set.

In this study, we explored the effectiveness of incorporating contextual metadata into Handwritten Text Recognition (HTR) models to enhance the digitization of medieval manuscripts. Utilizing the CATMuS Medieval dataset, which offers a rich variety of scripts, languages, and centuries, we compared the performance of Conformer models with and without contextual inputs, as well as training these models with auxiliary classification tasks. Our objective was to determine whether adding metadata such as Century, Script, and Language could improve model accuracy and robustness. We tested several configurations, including models with single and multiple contextual tokens, and evaluated them against both the baseline Conformer architecture and the original Kraken baselines. By doing so, we aimed to identify the most effective strategies for leveraging contextual information in HTR tasks.

Our results showed that the Conformer model using all contextual input tokens (Century, Script, Language) consistently outperformed other configurations, including the baseline models. This model achieved higher accuracy, particularly on the most challenging manuscripts, with an improvement of over 2 percentage points in some cases. Moreover, it exhibited the smallest variance in performance, indicating its robustness across different types of manuscripts. The use of multiple contextual tokens enabled the model to effectively learn and utilize diverse features, leading to better generalization. Interestingly, models with single contextual tokens did not perform as well and often fell short of the baseline, suggesting that a more comprehensive approach to metadata integration is necessary. Additionally, the auxiliary loss approach did not yield the expected improvements and frequently resulted in non-converging models, indicating the complexity of effectively balancing multiple training objectives.

While our approach demonstrated significant improvements, there are several areas for future exploration. The current approach relies on multi-hot encoding various categories without embedding these features into a learnable space beforehand. Approaches in natural language processing, such as [21], could potentially allow the model to approximate relationships between scripts and languages that are closely related, such as 'Caroline' and 'Humanistica' scripts. Secondly, the context token method appends information directly onto the image data fed into the encoder, a design choice motivated by the very lightweight FFN decoder which we deemed to be unlikely to effectively make use of the encoder features augmented with the raw context token. Combining a more powerful decoder, e.g. a pre-trained language model like in [22], and injecting metadata after the encoder is an avenue of future research. Such an architecture with a clear separation between the visual and linguistic model would presumably be beneficial for some types of semantic tokens, in particular language and genre, which we consider to be of more importance to the latter than the former.

Acknowledgments

Funded by the European Union (ERC, MiDRASH, Project No. 101071829). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

This work has received funding through the Investissements d'Avenir program of the Agence Nationale de la Recherche with reference ANR-21-ESRE-0005 (Biblissima+), and the DEFI Inria "Corpus et Outils pour les Langues de France".

References

[1] J.-B. Camps, N. Baumard, P.-C. Langlais, O. Morin, T. Clérice, J. Norindr, Make Love or War? Monitoring the Thematic Evolution of Medieval French Narratives, in: Computational Humanities Research (CHR 2023), CEUR-WS.org, 2023, pp. 734–756.

[2] S. Torres Aguilar, V. Jolivet, Handwritten Text Recognition for Documentary Medieval Manuscripts, Journal of Data Mining and Digital Humanities Historical Documents and automatic text recognition (2023). doi:10.46298/jdmdh.10484.

[3] D. Stutzmann, Fontenay Dataset. Original Charters From Fontenay before 1213, 2022.

[4] D. Stutzmann, S. T. Aguilar, P. Chaffenet, HOME-Alcar: Aligned and Annotated Cartularies, 2021.

[5] D. Stutzmann, Words as graphic and linguistic structures. Word spacing in Psalm 101 Domine exaudi orationem meam (eleventh-fifteenth centuries), in: Les Mots au Moyen Âge – Words in the Middle Ages, number 46 in Utrecht Studies in Medieval Literacy, Brepols, Turnhout, 2020, pp. 21–59. URL: 10.1484/M.USML-EB.5.120721.

[6] E. Gueville, D. J. Wrisley, Transcribing Medieval Manuscripts for Machine Learning, 2023. URL: https://shs.hal.science/halshs-03725166, working paper or preprint.

[7] A. Pinche, T. Clérice, A. Chagué, J.-B. Camps, M. Vlachou-Efstathiou, M. Gille Levenson, O. Brisville-Fertin, F. Boschetti, F. Fischer, M. Gervers, A. Boutreux, A. Manton, S. Gabay,

W. Haverals, M. Kestemont, C. Vandyck, P. O'Connor, CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts, in: DH2024, ADHO, Washington DC, United States, 2024. URL: https://inria.hal.science/hal-04346939.

[8] T. Clérice, A. Pinche, M. Vlachou-Efstathiou, A. Chagué, J.-B. Camps, M. Gille-Levenson, O. Brisville-Fertin, F. Fischer, M. Gervers, A. Boutreux, A. Manton, S. Gabay, P. O'Connor, W. Haverals, M. Kestemont, C. Vandyck, B. Kiessling, CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond, in: 2024 International Conference on Document Analysis and Recognition (ICDAR), Athens, Greece, 2024. URL: https://inria.hal.science/hal-04453952.

[9] R. Smith, An Overview of the Tesseract OCR Engine, in: Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02, ICDAR '07, IEEE Computer Society, USA, 2007, p. 629–633.

[10] B. Kiessling, Kraken - a Universal Text Recognizer for the Humanities, in: ADHO, Éd., Actes de Digital Humanities Conference, 2019.

[11] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 369–376.

[12] C. Tomoiaga, P. Feng, M. Salzmann, P. Jayet, Field Typing for Improved Recognition on Heterogeneous Handwritten Forms, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE Computer Society, 2019, pp. 487–493.

[13] A. K. Bhunia, S. Ghose, A. Kumar, P. N. Chowdhury, A. Sain, Y.-Z. Song, MetaHTR: Towards Writer-Adaptive Handwritten Text Recognition, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15825–15834. doi:10.1109/CVPR46437.2021.01557.

[14] A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, M. Stolz, Automatic Transcription of Handwritten Medieval Documents, in: 2009 15th International Conference on Virtual Systems and Multimedia, IEEE, 2009, pp. 137–142.

[15] A. Fischer, V. Frinken, A. Fornés, H. Bunke, Transcription Alignment of Latin Manuscripts using Hidden Markov Models, in: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, 2011, pp. 29–36.

[16] D. Stutzmann, J.-F. Moufflet, S. Hamel, La recherche en plein texte dans les sources manuscrites médiévales: enjeux et perspectives du projet HIMANIS pour l'édition électronique, Médiévales (2017) 67–96.

[17] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented Transformer for Speech Recognition, in: Proc. Interspeech 2020, 2020, pp. 5036–5040. doi:10.21437/Interspeech.2020-3015.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is All you Need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[19] D. Rekesh, N. Rao Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam, B. Ginsburg, Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition, arXiv e-prints (2023) arXiv:2305.05084. doi:10.

48550/arXiv.2305.05084. arXiv:2305.05084.
[20] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.
[21] J. Kim, R. K. Amplayo, K. Lee, S. Sung, M. Seo, S.-w. Hwang, Categorical Metadata Representation for Customized Text Classification, Transactions of the Association for Computational Linguistics 7 (2019) 201–215.
[22] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, F. Wei, TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 13094–13102.

A. Appendix

Table 8
Composition of the test dataset on CATMuS 1.5.0

Shelfmark	Language	Script	Type	Genre	Century	Lines	Characters
Paris, BnF, lat. 130	Latin	Caroline	prose	Treatises	12	198	13843
Paris, BnF, lat. 8001	Latin	Caroline	vers	Poetry	13	504	19550
Paris, BnF, lat. 7499	Latin	Caroline	prose	Treatises	10	**6086**	**168092**
Paris, BnF, fr. 1881	French	Cursiva	verse	Narratives	16	163	3507
Paris, BnF, fr. 604	French	Cursiva	verse	Narratives	15	343	10514
Paris, BnF, fr. 413	French	Cursiva	prose	Narratives	15	860	26952
Paris, BnF, lat. 14650	Latin	Cursiva	prose	Narratives	15	172	10752
Paris, Bibliothèque inter-universitaire de la Sorbonne, 193	Latin	Cursiva	prose	Treatises	14	669	34655
Paris, BnF, lat. 10996	Latin	Gothic Documentary Script	prose	Documents of practice	13	106	5548
Paris, BnF, esp. 368	Castilian	Humanistica	prose	Treatises	16	156	9092
Paris, BnF, ita. 481	Italian	Humanistica	prose	Narratives	14	502	18493
Florence, Biblioteca Medicea Laurenziana, Laur. Plut. 39.34	Latin	Humanistica	vers	Poetry	15	135	4268
Paris, BnF, Smith-Lesouëf 16	Latin	Humanistica	prose	Documents of practice	16	138	6415
Paris, BnF, esp. 36	Castilian	Hybrida	prose	Narratives	14	541	20043
Paris, BnF, lat. 17903	Latin	Praegothica	vers	Poetry	13	439	16228
Montpellier, Bibliothèque universitaire Historique de Médecine, H318	Latin	Praegothica	prose	Treatises	12	427	26773
Paris, BnF, Rés. YE-1325	French	Print	prose	Narratives	16	416	13957
Madrid, BNE, MSS. 3995	Castilian	Semihybrida	prose	Treatises	15	154	6178
Paris, BnF, fr. 2701	French	Semihybrida	prose	Treatises	15	273	13923
Paris, BnF, lat. 14137	Latin	Semitextualis	vers	Poetry	14	193	5706
Paris, BnF, fr. 574	French	Textualis	prose	Treatises	14	113	**2451**
Paris, BnF, fr. 13496	French	Textualis	prose	Narratives	13	159	4755
Paris, BnF, fr. 747	French	Textualis	prose	Narratives	13	**91**	5349
Paris, BnF, fr. 6447	French	Textualis	prose	Narratives	13	383	16310
Paris, BnF, fr. 23117	French	Textualis	prose	Narratives	13	736	24203
Paris, BnF, NAL 730	Latin	Textualis	prose	Treatises	14	284	14612
Vienna, ÖNB, 12.905	Middle Dutch	Textualis	prose	Treatises	14	1047	40465
Paris, BnF, esp. 65	Navarrese	Textualis	prose	Treatises	14	736	19932
Paris, BnF, ita. 783	Venetian	Textualis	prose	Narratives	14	147	7361