The LAM Dataset: A Novel Benchmark for Line-Level Handwritten Text Recognition

Silvia Cascianelli[1], Vittorio Pippi[1], Martin Maarand[2], Marcella Cornia[1], Lorenzo Baraldi[1], Christopher Kermorvant[2], Rita Cucchiara[1]
[1]University of Modena and Reggio Emilia, Modena, Italy [2]TEKLIA, Paris, France
Email: [1]{name.surname}@unimore.it, [2]{surname}@teklia.com

***Abstract*—Handwritten Text Recognition (HTR) is an open problem at the intersection of Computer Vision and Natural Language Processing. The main challenges, when dealing with historical manuscripts, are due to the preservation of the paper support, the variability of the handwriting – even of the same author over a wide time-span – and the scarcity of data from ancient, poorly represented languages. With the aim of fostering the research on this topic, in this paper we present the Ludovico Antonio Muratori (LAM) dataset, a large line-level HTR dataset of Italian ancient manuscripts edited by a single author over 60 years. The dataset comes in two configurations: a basic splitting and a date-based splitting which takes into account the age of the author. The first setting is intended to study HTR on ancient documents in Italian, while the second focuses on the ability of HTR systems to recognize text written by the same writer in time periods for which training data are not available. For both configurations, we analyze quantitative and qualitative characteristics, also with respect to other line-level HTR benchmarks, and present the recognition performance of state-of-the-art HTR architectures. The dataset is available for download at https://aimagelab.ing.unimore.it/go/lam.**

I. Introduction

Handwritten Text Recognition (HTR) has been studied for decades [1], [2], [3], thanks to its importance in terms of practical applications (ranging from public administration to industrial processes automation and digital humanities) and for its multimodal and sequential nature, that is common to other pattern recognition tasks. Despite the encouraging results achieved by the recent literature, and especially by deep learning-based models [4], [5], [6], [7], HTR is still far from being considered a solved task.

When performing HTR on historical manuscripts [8], [9], [10], [11], [12], [13], there are additional challenges which need to be taken into account, which are both visual and linguistic issues. From the visual point of view, the digitized images of an historical manuscript exhibit several artifacts: both the paper support and the ink can be deteriorated, and there can be stains, scratches, bleed-through or faded ink. Further, the language used is typically peculiar to the historic period and geographical area in which the manuscript was edited. This often prevents exploiting a pre-trained language model in modern English language and represents a challenge from the textual and linguistic point of view. Designing effective strategies for the challenges mentioned above requires rich data collections, manually annotated.

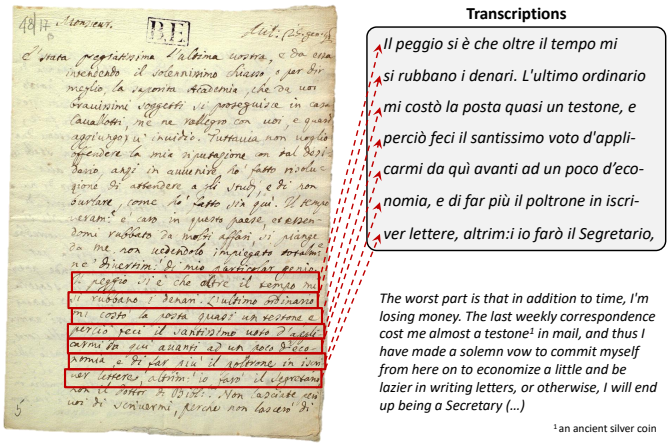


Fig. 1. The LAM dataset features lines from letters by the Italian historian L. A. Muratori. To the best of our knowledge, it is the largest dataset for line-level HTR.

In this paper we contribute to the research on handwriting recognition by presenting a novel and large dataset for HTR on historical manuscripts. The dataset is obtained from letters handwritten by the Italian historian Ludovico Antonio Muratori (1672-1750), which are preserved at the Estense Library of Modena (Italy). We selected letters written personally by Muratori (discarding those written by his collaborators, thanks to a precise evaluation by experts), and entirely in ancient Italian[1]. While all letters have been written by a single author, they cover a large time-span of around 60 years – during which the author's handwriting and paper support varied. This makes the dataset suitable for dealing with the change of handwriting style over time, in addition to being a pure HTR dataset. The annotation of the dataset has been performed manually, and it has been double-checked by two experts, who provided diplomatic transcriptions at line-level. This level of annotation granularity has been chosen as it is a good trade-off between word-level and paragraph-level in terms of required time, cost, and amount of supervision and because it is common in HTR research. Overall, the proposed dataset contains 25,823 lines, which, to the best of our knowledge, makes it *the largest line-level HTR dataset to date*.

In the reminder of the paper, we present the LAM dataset,

[1]Muratori, indeed, wrote parts of letters in Spanish, French, and Latin, depending on the correspondent.

named after the initials of Ludovico Antonio Muratori, and provide an overview of its main features by comparing it with existing proposals. Moreover, we perform a quantitative analysis of the performance achieved by different HTR approaches, including both state-of-the-art models and tools, on the dataset with the aim of providing the community with baselines and insights for developing novel architectures for HTR on historical data.

II. Related Work

Depending on the choice of the textual unit, HTR can be performed at different granularity levels, ranging from character-level [14], [15] (which is particularly suited for idiomatic languages) to page-level [16], [17], [18], [19], [20]. Line-level HTR is a popular trade-off and one of the most studied variants, especially for non-idiomatic languages. In fact, line-level HTR not only is a standalone variant of the task, but is also often integrated in paragraph-level or page-level HTR systems [6], [16], [17]. For this reason, the LAM dataset has been designed to fit a line-level protocol.

The first approaches to HTR entailed the use of Hidden Markov Models (HMMs), in combination with Gaussian Mixture Models or fully-connected networks for representing the visual input, and n-gram based language models for predicting the textual output [21], [22]. In following works, the visual input representation has been performed by using Multi-Dimensional Long Short-Term Memory networks (MD-LSTMs), and the Connectionist Temporal Classifier (CTC) decoding strategy has been introduced to produce the transcription [18], [23], [24], as proposed in [3]. Alternatively to MD-LSTMs, CNNs can be used, in combination with one-dimensional LSTMs, to encode the text image [4], [5]. This strategy later became a popular choice [25], [26], [27], [28], [29], [30] since it allows reaching comparable or superior results to MD-LSTM-based approaches, while being faster to train. Other approaches have also been investigated to avoid the usage of Recurrent Neural Networks (RNNs). For example, in [31], it is proposed an hybrid approach combining convolutional layers and time-delay neural layers [32] for input representation, with an HMM for output prediction. Further, in [6], [33], Fully Convolutional Networks (FCNs) are proposed for HTR. To simulate the dependency modeling provided by LSTM cells, FCN are combined with GateBlocks layers [34], which implement a selecting mechanism similar to that of LSTM cells. Each gate is made of Depth-wise Separable Convolutions [35] to reduce the number of parameters and speed up the training process. A recent research line has proposed to apply the sequence-to-sequence paradigm to HTR [36], [37], where the text image is encoded via convolutional and recurrent layers, and the transcription is generated by a RNN-based decoder. As training objective, the CTC loss commonly used in HTR can be combined with the cross-entropy loss. As a special case of the sequence-to-sequence paradigm, some works apply Transformers [38] as encoders or decoders [39], [40], motivated by the success of such architecture in machine translation and language understandings and other vision-and-language tasks [41], [42], [43], [44]. In HTR, this strategy is effective when sufficient training data is available. For this reason, some works employ synthetic data during a pre-training stage [7], [45], [46]. Finally, to increase performance, many approaches integrate an explicit language model. This strategy is as effective as the language in the dataset is regular and well-represented, which is not often the case for historical datasets.

TABLE I
Characteristics of line-level benchmark datasets.

	Lines	Lexicon	Period	Language	Authors
IAM [47]	10,373	9,749	Modern	English	Many
RIMES [48]	12,111	8,760	Modern	French	Many
Washington[49]	656	1,471	1755	English	Two
Saint Gall [50]	1,410	5,436	ca 890-900	Latin	One
Esp. Index [51]	1,563	1,725	1491-1495	Catalan	One
Leopardi [12]	2,459	5,067	1818-1832	Italian	One
Parzival [49]	4,477	4,934	ca 1265-1300	German	Two
Esp. Licenses [51]	5,447	3,465	1616-1619	Catalan	One
ICFHR14 [52]	11,473	9,716	ca 1760-1832	English	Many
ICFHR16 [53]	10,550	8,120	1470-1805	German	Many
ICFHR18 [54]	14,803	23,198	Mixed[1]	German, Italian	Many
Rodrigo [55]	20,357	17,300	1545	Spanish	One
Germana [56]	20,529	27,100	1891	Spanish[2]	One
LAM (Ours)	25,823	23,428	1691-1750	Italian	One

1 - Both Medieval and Modern.
2 - A small number of lines are in different languages.

Existing Benchmark Datasets. Designing and developing effective HTR solutions requires the availability of large data collections, which should capture both the visual variability of the task and represent different languages. In the following, we focus on line-level dataset of western-characters, since these are more closely related to our proposed dataset. The main characteristics of those datasets are also reported in Table I.

Commonly-used benchmarks for line-level HTR include the IAM [47] and the RIMES [48] datasets, both containing lines in modern languages (English and French, respectively) and written by multiple authors on regular paper support. The language used is somewhat constrained, since the writers have been carefully instructed on what to write: copying English sentences from the Lancaster-Oslo/Berge corpus [57] in IAM, and following a template and a script for writing customer service-themed letters in RIMES.

As for HTR on historical manuscripts, many datasets have been released to explore different perspectives of the task. Among those, the most used are the ICFHR14 [52], ICFHR16 [53], and the ICFHR18 [54] datasets, all prepared for HTR challenges at the International Conference on Frontiers of Handwriting Recognition (ICFHR). The aim of ICFHR14 is to explore HTR on historical data rather than modern ones. Therefore, the dataset contains lines form the Bentham Papers collection [58], handwritten in English by few authors, mainly the philosopher Jeremy Bentham and his collaborators. The ICFHR16 was initially intended to study HTR on a language that is structurally different from English, and thus it contains lines from the Ratsprotokolle collection, handwritten in German by multiple writers in over

three centuries. Finally, the ICFHR18 dataset was designed to investigate the minimum amount of training data required to correctly transcribe an entire historical document. For this reason, the dataset contains documents from many different collections and time periods, it is written in different languages (German and Italian), and its test set is divided in document-specific sets. Moreover, it is worth mentioning the Rodrigo [55] and Germana [56] datasets. These are large line-level datasets obtained from two different Spanish books and written by a single author in a short time-span.

There are also other datasets containing historical manuscripts, which are of much smaller size and thus can be used to explore HTR in the case of limited training data and specific domain. Examples of such datasets are the George Washington dataset [49], containing English letters by George Washington (and few parts by a collaborator), and the Parzival dataset [50], containing a Medieval German poem handwritten by two writers. Usually, datasets of this kind feature documents handwritten by a single author in a relatively limited time-span, during which the handwriting does not change significantly. Some examples are the Saint Gall dataset [59], with lines from a Medieval Latin manuscript, the Esposalles Index and Esposalles Licenses datasets [51], with lines from Catalan marriages registers, and the Leopardi dataset [12], with Italian letters by the writer Giacomo Leopardi.

III. The LAM Dataset

In this section, we analyze the main characteristics of the proposed dataset. It comes with different splittings to allow performing classical HTR, on a splitting we refer to as *basic split*, and time-dependent HTR, in a setting we refer to as *date-based setting*. The main characteristics of these splittings are reported in Table II.

A. Data Collection and Preparation

The documents used for the LAM dataset come from the digitized L. A. Muratori collection preserved at the Estense Digital Library. The collection contains drafts, papers, notes, and letters handwritten by the Italian historian and his collaborators. Some of these documents, or parts of those, are written in languages different from Italian, which include Latin, French, and Spanish.

For collecting the dataset, we prepared an ad-hoc on-line annotation tool. We preferred not to use available commercial tools to obtain simplicity of use by non-experts, to keep the data in house before the release of this dataset, and to favor crowd-sourcing since the tool does not require any license or subscription to be used. Further details on the developed platform can be found in the supplementary material.

Two experts were involved in the data preparation. First, they selected from the considered collection of digitized documents, only autograph letters by Muratori in Italian, for a total of 1,171 pages from 72 files edited in a time-span of 60 years. Then, they annotated the letters at line-level, by providing the bounding box of each line and its diplomatic transcription. In the transcription process, stroke-out text, words that are illegible due to stains and scratches, and special symbols not representable in Unicode have been replaced with "#".

TABLE II
LAM dataset splits statistics. The charset size is calculated on the Training and Validation splits.

	Total	Training	Validation	Test	Charset
Basic split	25,823	19,830	2,470	3,523	89
Date-based setting (leave-decade-out)					
1690-1700	25,183	17,205	1,911	6,067	87
1700-1710	22,392	17,205	1,911	3,276	84
1710-1720	21,066	17,205	1,911	1,950	83
1720-1730	25,158	17,205	1,911	6,042	86
1730-1740	22,974	17,205	1,911	3,858	85
1740-1750	23,106	17,205	1,911	3,990	84
Date-based setting (decade-vs-decade)					
1690-1700	25,183	5,460	607	19,116	80
1700-1710	25,183	2,948	328	21,907	80
1710-1720	25,183	1,755	195	23,233	77
1720-1730	25,183	5,437	605	19,141	81
1730-1740	25,183	3,472	386	21,325	83
1740-1750	25,183	3,591	399	21,193	81

Note that each considered file contains letters to a different correspondent, which was either a family member, a friend, a professional or an acquaintance of different social extraction and cultural level, to whom Muratori wrote about many different topics. This results in a rich and varied language. The annotation process and subsequent double-checking took approximately one year and, to the best of our knowledge, lead to the largest dataset for line-level HTR to date.

B. General Characteristics

The LAM dataset contains a total of 25,823 lines, with a lexicon of over 23,000 unique words (see Table I). Other datasets of comparable size feature a rich lexicon as well, especially those containing text in different languages (*e.g.* the ICFHR18 and Germana datasets). In the case of the LAM dataset, the richness of the lexicon is due to the fact that, in his letters, Muratori wrote about different topics, mentioned many different proper nouns (of people and places), and used various forms of abbreviations for names, titles, and salutations, which was common in his time to save paper when writing.

As for the visual characteristics, the LAM dataset features all the typical time-related challenges of historical manuscripts. In particular, the paper support used varies considerably from page to page (both in terms of color and texture), and there are pages with humidity stains, creases, scratches, and holes. Also the ink used makes the dataset challenging since there are pages with faded or bled trough ink, stains, discolorations, and corrosions. Some pages from the LAM dataset exemplifying the aforementioned challenges are shown in the supplementary material.

Further characteristics of the LAM dataset are analyzed in Fig. 2 in comparison with other commonly used benchmark datasets, both modern and historical, and with a smaller historical dataset in Italian (the Leopardi dataset). Compared to other datasets, the lines in LAM have smaller and more regular height, while the line images width has a more clear tendency to bimodality. This is due to the fact that, depending on the content and the addressee, some pages are written in double-column. Moreover, to save paper, the author commonly

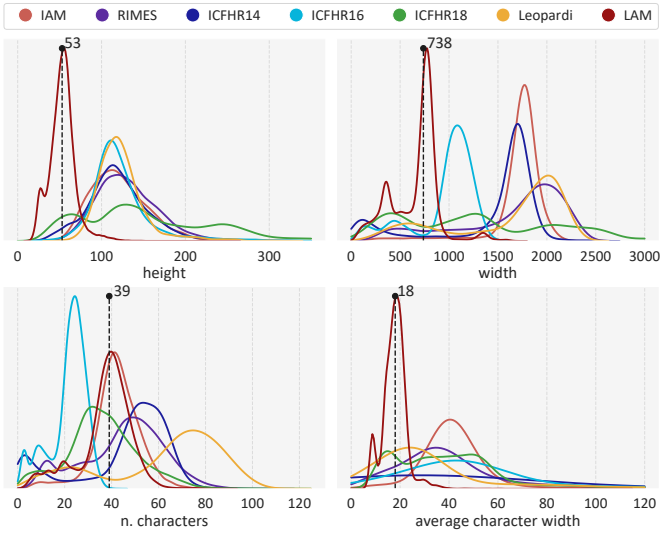


Fig. 2. Lines height and width distribution in the LAM dataset compared to other benchmark datasets (top). Number of characters per line distribution in the LAM dataset compared to other popular benchmark datasets (bottom left). Average characters pixel width distribution in the LAM dataset compared to other benchmark datasets (bottom right). Best seen in color.

exploited the entire text column width, which also explains the clearer regularity in the width of the characters compared to other datasets. Also in terms of the average number of characters per line LAM shows regularity, having the majority of lines with 39 characters (similar to IAM).

To use the LAM dataset for classic HTR, we provide a **basic split** consisting of 19,830 lines for training, 2,470 for validation, and 3,523 for test. The lines have been collected from different portions of the pages in each of the 72 considered files, of 80%, 10%, and 10%, respectively. This splitting is intended for exploring HTR on images featuring the typical challenges of historical manuscripts and a rich and underrepresented language such as ancient Italian.

C. Date-based Setting

As mentioned above, the documents from which we collected the LAM dataset cover roughly 60 years. For most of the letters in the files the date in which they have been written is clearly indicated. Therefore, according to this information, we were able to separate them into six groups reflecting the decade. The idea behind this date-based setting is to explore the effect of the availability of handwritten samples from an author in different time periods over the recognition of his/her text in an unseen time period. In a wide time-span as that considered in the LAM dataset, the handwriting of the author is likely to change. In this respect, a t-SNE analysis of the lines in the date-based setting and examples of pages from the six splits can be found in the supplementary material.

After discarding pages with no date indication (27 out of 1,171 pages), we built two setups that can be used to perform HTR of the same author, conditioned on time. In the first setup, referred to as **leave-decade-out**, the test set contains all the lines from the pages of a certain decade, while the training and validation sets contain a proportional amount of lines from the pages of the other decades (90% and 10%). Note that, for fair comparison and data balancing, we include the same amount of lines in the training set and the validation set of each split. In the second setup, referred to as **decade-vs-decade**, all the lines from a decade of choice are used for training and validation, and all the lines from each other decade separately are used for test. The size of the subsets and the charset in each date-based split is reported in Table II.

IV. Experimental Evaluation

In this section, we report an experimental analysis of the performance of popular state-of-the-art models and toolkits, both on the basic split of the LAM dataset and on the date-based setting. The performance are reported in terms of Character Error Rate (CER) and Word Error Rate (WER). As customary in HTR, to calculate the CER and the WER on the entire test set, we first compute the Edit Distance (at character level for the CER and word level for the WER) between each predicted sentence and the corresponding ground truth. This is the number of substitutions, deletions, and insertions that have to be applied to the predicted sentence to obtain the ground truth. Then, we sum up the distances of all samples, divide by the sum of the ground truth lengths and multiply by 100.

A. Considered HTR Approaches

We consider models following different kinds of architectures for HTR in order to give insights on the possibly more promising strategy to be applied on the LAM dataset and guide future research. When available, we used the official implementation and weights provided by the authors, while in the other cases, we used our best implementation. All the models have been trained by following the training protocol described in the original paper. Note that data augmentation is not performed in any of the considered approaches to better highlight the effect of the size of the LAM dataset and the data variability it captures. For methods requiring an explicit charset, this has been obtained from the training and validation subsets of the basic split, and of each of the splittings in the date-based setting (see Table II).

1) Convolutional-Recurrent Paradigm: Combining CNNs and RNNs for HTR has been the standard choice for years. In this analysis we consider models featuring 1D-LSTMs, since these have been proven to be comparable or superior to MDLSTMs [3] while being much faster to train [5]. In particular, we test our implementation of the **1D-LSTM** proposed in [5], which consists of a stack of five convolutional blocks and five Bidirectional Long Short-Term Memory network (BLSTM) layers. We also consider the approach proposed in [4] (referred to as **CRNN** in the following), which has a deeper convolutional component but fewer recurrent layers compared to 1D-LSTM. Specifically, in this variant there are seven convolutional blocks, two of whom contain rectangular max-pooling layers to better maintain the aspect ratio of the text lines, and two BLSTM layers. For both the 1D-LSTM and CRNN approaches, we additionally consider variants containing Deformable Convolutions (DefConvs) [60], as proposed

in [12], [29]. Finally, we include in the analysis the default model available from the popular HTR toolkit **PyLaia** [61], which has four convolutional layers and three BLSTM layers.

2) Sequence-to-Sequence Paradigm with Transformers: As a representative of the sequence-to-sequence paradigm, we explore Transformer-based approaches, which are more data-demanding than classical RNN-based solutions. Therefore, by considering these kind of architectures, we can investigate whether the size of the LAM dataset allows effectively training large HTR models. In this respect, we consider the strategy proposed in [7] (referred to as **Transformer** in the following). This architecture exploits a ResNet-101 trained from scratch and a Transformer encoder and decoder [38] with reduced parameters. The ResNet produces a feature map that is then flattened and used as input embeddings for the Transformer architecture. Moreover, we consider the Base version of the **TrOCR** model proposed in [46], which employs Transformer-like architectures both for image representation [62], [63] and text generation [38], [41], and exploits large-scale pre-training, both on typewritten and handwritten lines, before being fine-tuned on the dataset of interest.

3) Fully-Convolutional Paradigm: Recent approaches to HTR come with FCNs, establishing state-of-the-art results. In this work, we consider the Gated Fully Convolutional Network **(GFCN)** [33], which preprocesses the input image with four convolutional layers and then passes the output through five GateBlocks layers [34]. We also consider three different variants of the deeper model **OrigamiNet** [6], containing 12, 18, and 24 GateBlocks layers, respectively. Moreover, we include in the analysis the implementation of the approach proposed in [31] available in the **Kaldi** toolkit. This model is composed of six convolutional layers and three time-delay neural layers, followed by an HMM for text recognition. The architecture training is divided into two phases. First, a "flat start" model is trained on images and the corresponding transcriptions. Then, the trained "flat start" model is used to create alignments for training a second model, which is the only one used at inference time. Finally, a 3-gram byte pair encoding language model is applied to improve the decoding.

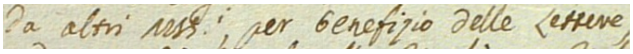
B. Evaluation Results

LAM Basic Split. The results obtained by the models included in the analysis on the basic split are reported in Table III. In this setting, the models following the convolutional-recurrent paradigm obtain a CER on average below 4% and a WER below 12%. The performance improvement given by the use of DefConvs indicates the importance of a strong image encoder in this dataset. Among the approaches following the fully-convolutional paradigm, only OrigamiNet performs on par with the convolutional-recurrent approaches, and even performs best in its variant containing 24 GateBlocks. This can be traced back to the high number of GateBlocks that this model contains, allowing to better model the context compared to GFCN and Kaldi, confirming the importance of the image representation for this dataset. Another observation comes from the relatively poor performance of the Transformer model. This implements an intrinsic language model that is challenged by the heavy use of rare words and abbreviations in this dataset. The TrOCR model, which features both a strong image representation component and an intrinsic language model pre-trained on a large amount of image-text pairs, reaches error rates that are comparable with those of the convolutional-recurrent models after being fine-tuned for 30 epochs on the LAM dataset.

TABLE III
RESULTS ON THE TEST SETS OF THE LAM BASIC SPLIT, AND ON THE IAM AND THE ICFHR14 DATASETS. THE * MARKER INDICATES OUR RE-IMPLEMENTATIONS.

		IAM		ICFHR14		LAM	
Method	**#Params (M)**	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**
HTR Toolkits							
PyLaia [61]	4.8	9.8	32.3	5.1	17.5	4.7	16.5
Kaldi [31]	15.0	7.2	25.0	3.7	14.2	4.7	13.4
HTR Models							
1D-LSTM [5]	-	8.3	24.9	-	-	-	-
1D-LSTM [5]*	9.6	7.7	26.3	4.8	15.3	3.7	12.3
1D-LSTM (w/ DefConv) [29]	9.6	7.5	26.9	3.6	14.3	3.5	11.6
CRNN [4]*	18.2	7.8	27.8	3.9	15.3	3.8	12.9
CRNN (w/ DefConv) [29]	18.5	6.8	24.7	3.6	13.9	3.3	11.3
Transformer [7]*	54.7	-	-	-	-	10.2	22.0
TrOCR [46]	385.0	3.4	-	-	-	-	-
TrOCR [46]*	385.0	7.3	37.5	3.5	11.5	3.6	11.6
GFCN [33]	1.4	8.0	28.6	-	-	-	-
GFCN [33]*	1.4	8.0	28.6	-	-	5.2	18.5
OrigamiNet$_{12}$ [6]	39.0	5.3	-	-	-	-	-
OrigamiNet$_{12}$ [6]*	39.0	6.0	22.3	3.6	14.7	3.1	11.2
OrigamiNet$_{18}$ [6]	77.1	4.8	-	-	-	-	-
OrigamiNet$_{18}$ [6]*	77.1	6.6	24.2	4.0	15.4	3.1	11.0
OrigamiNet$_{24}$ [6]	115.3	4.8	-	-	-	-	-
OrigamiNet$_{24}$ [6]*	115.3	6.5	23.9	5.9	21.3	3.0	11.0

Ground-truth	Um.mo Div.mo ed Osseq.mo Ser.re e Sud.o
CRNN (w/ DefConv)	Vo.mo div.mo ed Ossegl.mo Ser.r e Sud.o
TrOCR	Um.mo Div.mo ed ossequ.mo Ser.re e Sud.o
OrigamiNet$_{24}$	Um.mo Di.o ed Ossegl.mo Ser.r e Sud.e



Ground-truth	da altri # per benefizio delle Lettere,
CRNN (w/ DefConv)	da altri MSS.i, per benefizio delle Lettere,
TrOCR	da altri Mess.i, per benefizio delle Lettere
OrigamiNet$_{24}$	da altri MS#.i, per benefizio delle Lettere

Fig. 3. Qualitative results of the best performing models on example challenging lines from the LAM dataset.

As a further comparison between the LAM dataset and other existing benchmarks, we evaluate the performance of the considered models also on IAM and ICFHR14. A performance drop can be noticed for all models, especially bigger ones, some of which did not converge in some cases (*i.e.* Transformer and GFCN). This indicates that the large amount of training data provided by the LAM dataset can contribute to enable the development of effective models for HTR, similar to what has been done for other vision-and-language tasks.

Finally, we report the qualitative results of the best performing models following the three considered HTR paradigms (CRNN with DefConv, TrOCR, and OrigamiNet$_{24}$) on challenging lines of the LAM dataset in Fig. 3. Additional examples can be found in the supplementary material.

LAM Date-based Setting. As for the leave-decade-out setup

TABLE IV
RESULTS ON THE LEAVE-DECADE-OUT SETUP OF THE DATE-BASED SETTING. THE * MARKER INDICATES OUR RE-IMPLEMENTATIONS.

	1690-1700		1700-1710		1710-1720		1720-1730		1730-1740		1740-1750		Average	
Method	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**
HTR Toolkits														
PyLaia [61]	6.0	23.3	3.7	13.7	3.1	11.2	3.0	11.5	4.8	16.1	3.9	14.5	4.0	15.1
Kaldi [31]	4.9	19.1	3.0	10.4	2.7	9.7	2.5	9.4	4.5	13.1	3.2	11.4	3.5	12.2
HTR Models														
1D-LSTM [5]*	5.3	20.9	3.7	12.6	2.8	9.1	2.7	9.0	4.1	14.3	3.6	11.8	3.7	13.0
1D-LSTM (w/ DefConv) [29]	5.0	19.8	3.6	12.1	2.5	8.3	3.1	10.1	3.8	12.3	3.6	12.2	3.6	12.5
CRNN [4]*	5.0	20.1	3.5	12.1	2.6	8.8	3.1	10.4	4.3	14.0	3.8	12.7	3.7	13.0
CRNN (w/ DefConv) [29]	4.7	19.0	3.3	11.1	2.2	7.6	2.4	8.3	3.7	12.2	3.4	11.1	3.3	11.6
Transformer [7]*	15.2	37.2	18.6	36.5	14.6	28.5	10.0	20.1	19.5	35.7	11.9	25.2	15.0	30.5
GFCN [33]	5.1	18.5	7.4	23.1	3.0	10.8	4.2	14.6	5.5	17.7	4.2	15.4	4.9	16.7
OrigamiNet$_{12}$ [6]	4.6	18.9	2.8	10.3	2.2	8.0	2.2	8.3	3.4	11.4	3.0	11.8	3.0	11.5
OrigamiNet$_{18}$ [6]	4.5	18.7	2.8	10.3	2.2	8.1	2.3	8.9	3.5	11.8	2.3	8.9	2.9	11.1
OrigamiNet$_{24}$ [6]	4.9	20.4	2.9	10.6	2.3	8.2	2.2	8.4	3.3	10.9	3.1	11.6	3.1	11.7

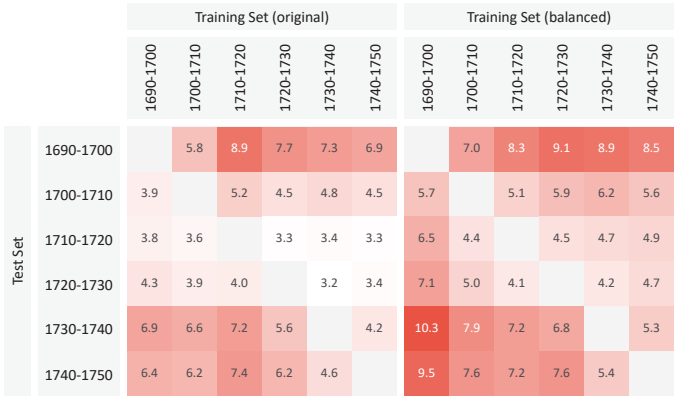


Fig. 4. Results of OrigamiNet$_{18}$ in the decade-vs-decade setup of the date-based setting, both in the scenario in which all the training samples available for each decade are used (left) and the balanced scenario (right).

of the date-based setting, we report the results of the considered approaches in Table IV. For all the models, the worst performance is obtained on the fifth-decade split (1730-1740). The first and second splits are challenging as well, considering the performance of GFCN and Transformer on the second decade, and of all the other approaches on the first decade. The splits having the third and fourth decade as test set are instead easier to recognize. In fact, the errors of the considered approaches on these splits are even lower than what obtained on the basic split. Arguably, this is due to a more homogeneous and clear handwriting of the author in his middle age. Additional to a decade-specific analysis, to express the overall performance of HTR models to recognize text over time, we propose to use the average CER and WER on the six splits. According to these scores, the best-performing model in this setting is OrigamiNet in its variant with 18 GateBlocks.

To further explore the challenges posed by the date-based setting, we consider OrigamiNet$_{18}$ in the decade-vs-decade setup. The CER values obtained in this experiment are reported in Fig. 4 (the WER scores are reported in the supplementary material). Overall, the first and the last two decades are the most challenging to recognize, while the text produced in the author's middle age is easier to recognize. The results reported in the table also highlight the difficulty in transcribing documents written at an early age when training on those written at a late age and vice versa, which is a challenge posed by the date-based setting. Moreover, to assess whether the difference in performance can be attributed to the difference in the number of samples available for each decade, we repeat the above analysis by using training sets of equal size (artificially balanced by randomly sampling an equal number of lines for each decade) and the same test sets as in Table II (further details are given in the supplementary material). The results of this analysis are reported in Fig. 4. Despite the expected numerical variations in the specific CER values, the same observations made in the case of the released setup apply also in this case of balanced setup, thus allowing imputing the challenges emerged to the characteristics of the data rather than to the training set size.

V. CONCLUSION

In this work, we presented the LAM dataset for line-level HTR on historical manuscripts, containing more than 25,000 lines. The dataset features letters written in Italian by a single author over around 60 years, which makes it suitable not only for research on HTR, but also on handwriting recognition over time. To this end, the dataset comes with different splits, reflecting the decade in which the letters have been written. Quantitative and qualitative analyses of the dataset, both of its characteristics and performance achievable with commonly used HTR approaches, highlight the challenges posed by the LAM dataset, which we hope can make it a valuable contribution towards the development of effective solutions to HTR on historical documents. As a further development of this work, the LAM dataset could be enriched with word-level annotations, thus increasing the level of supervision and making it suitable also for the keyword spotting task on historical manuscripts.

ACKNOWLEDGMENT

This work was supported by the "AI for Digital Humanities" project , funded by "Fondazione di Modena", by the "DHMoRe Lab" project , funded by "Regione Emilia Romagna", and by the "Artificial Intelligence for Cultural Heritage" project, co-funded by the Italian Ministry of Foreign Affairs and International Cooperation. The authors thank Dr. Maria Ludovica Piazzi, Dr. Rosiana Schiuma, and the Estense Digital Library for the contribution and support provided in preparing the dataset.

REFERENCES

[1] U.-V. Marti and H. Bunke, "Handwritten sentence recognition," in *Proceedings of the International Conference on Pattern Recognition*, 2000.

[2] E. Krevat and E. Cuzzillo, "Improving off-line handwritten character recognition with hidden markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, 2006.

[3] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2009.

[4] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[5] J. Puigcerver, "Are multidimensional recurrent layers really necessary for handwritten text recognition?" in *Proceedings of the International Conference on Document Analysis and Recognition*, 2017.

[6] M. Yousef and T. E. Bishop, "OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recognition by learning to unfold," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[7] L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, "Pay attention to what you read: Non-recurrent handwritten text-line recognition," *arXiv preprint arXiv:2005.13044*, 2020.

[8] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal on Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 139–152, 2007.

[9] F. Bolelli, G. Borghi, and C. Grana, "XDOCS: An application to index historical documents," in *Proceedings of the Italian Research Conference on Digital Libraries*, 2018.

[10] M. Bouillon, R. Ingold, and M. Liwicki, "Grayification: a meaningful grayscale conversion to improve handwritten historical documents analysis," *Pattern Recognition Letters*, vol. 121, pp. 46–51, 2019.

[11] A. Santoro and A. Marcelli, "Using keyword spotting systems as tools for the transcription of historical handwritten documents: Models and procedures for performance evaluation," *Pattern Recognition Letters*, vol. 131, pp. 329–335, 2020.

[12] S. Cascianelli, M. Cornia, L. Baraldi, M. L. Piazzi, R. Schiuma, and R. Cucchiara, "Learning to Read L'Infinito: Handwritten Text Recognition with Synthetic Training Data," in *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, 2021.

[13] J. C. Aradillas, J. J. Murillo-Fuentes, and P. M. Olmos, "Boosting offline handwritten text recognition in historical documents with few labeled lines," *IEEE Access*, 2021.

[14] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015.

[15] N. D. Cilia, C. De Stefano, F. Fontanella, and A. S. di Freca, "A ranking-based feature selection approach for handwritten character recognition," *Pattern Recognition Letters*, vol. 121, pp. 77–86, 2019.

[16] B. Moysset, C. Kermorvant, and C. Wolf, "Full-page text recognition: Learning where to start and when to stop," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2017.

[17] T. Bluche, J. Louradour, and R. Messina, "Scan, Attend and Read: End-to-End Handwritten Paragraph Recognition with MDLSTM Attention," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2017.

[18] T. Bluche, "Joint line segmentation and transcription for end-to-end handwritten paragraph recognition," in *Advances in Neural Information Processing Systems*, 2016.

[19] C. Wigington, C. Tensmeyer, B. Davis, W. Barrett, B. Price, and S. Cohen, "Start, Follow, Read: End-to-End Full-Page Handwriting Recognition," in *Proceedings of the European Conference on Computer Vision*, 2018.

[20] T. Clanuwat, A. Lamb, and A. Kitamoto, "KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2019.

[21] A. H. Toselli, A. Juan, J. González, I. Salvador, E. Vidal, F. Casacuberta, D. Keysers, and H. Ney, "Integrated handwriting recognition and interpretation using finite-state models," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 04, pp. 519–539, 2004.

[22] A. H. Toselli and E. Vidal, "Handwritten text recognition results on the Bentham collection with improved classical n-gram-HMM methods," in *Proceedings of the International Workshop on Historical Document Imaging and Processing*, 2015.

[23] B. Moysset and R. Messina, "Are 2D-LSTM really dead for offline text recognition?" *International Journal on Document Analysis and Recognition*, vol. 22, no. 3, pp. 193–208, 2019.

[24] G. M. de Buy Wenniger, L. Schomaker, and A. Way, "No padding please: Efficient neural handwriting recognition," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2019.

[25] V. Pham, T. Bluche, C. Kermorvant, and J. Louradour, "Dropout improves recurrent neural networks for handwriting recognition," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 2014.

[26] P. Voigtlaender, P. Doetsch, and H. Ney, "Handwriting recognition with large multidimensional long short-term memory recurrent neural networks," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 2016.

[27] T. Bluche and R. Messina, "Gated convolutional recurrent neural networks for multilingual handwriting recognition," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2017.

[28] A. Chowdhury and L. Vig, "An efficient end-to-end neural model for handwritten text recognition," in *Proceedings of the British Machine Vision Conference*, 2018.

[29] I. Cojocaru, S. Cascianelli, L. Baraldi, M. Corsini, and R. Cucchiara, "Watch your strokes: Improving handwritten text recognition with deformable convolutions," in *Proceedings of the International Conference on Pattern Recognition*, 2020.

[30] S. Cascianelli, M. Cornia, L. Baraldi, and R. Cucchiara, "Boosting modern and historical handwritten text recognition with deformable convolutions," *International Journal on Document Analysis and Recognition*, pp. 1–11, 2022.

[31] A. Arora, C. C. Chang, B. Rekabdar, B. BabaAli, D. Povey, D. Etter, D. Raj, H. Hadian, J. Trmal, P. Garcia *et al.*, "Using ASR methods for OCR," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2019.

[32] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[33] D. Coquenet, C. Chatelain, and T. Paquet, "Recurrence-free unconstrained handwritten text recognition using gated fully convolutional network," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 2020.

[34] M. Yousef, K. F. Hussain, and U. S. Mohammed, "Accurate, Data-Efficient, Unconstrained Text Recognition with Convolutional Neural Networks," *Pattern Recognition*, vol. 108, p. 107482, 2020.

[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[36] J. Sueiras, V. Ruiz, A. Sanchez, and J. F. Velez, "Offline continuous handwriting recognition using sequence to sequence neural networks," *Neurocomputing*, vol. 289, pp. 119–128, 2018.

[37] J. Michael, R. Labahn, T. Grüning, and J. Zöllner, "Evaluating sequence-to-sequence models for handwritten text recognition," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2019.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[39] C. Wick, J. Zöllner, and T. Grüning, "Transformer for Handwritten Text Recognition Using Bidirectional Post-decoding," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2021.

[40] D. H. Diaz, S. Qin, R. Ingle, Y. Fujii, and A. Bissacco, "Rethinking Text Line Recognition Models," *arXiv preprint arXiv:2104.07787*, 2021.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[43] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-Shot Text-to-Image Generation," *arXiv preprint arXiv:2102.12092*, 2021.

[44] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-based Image Captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[45] C. Wick, J. Zöllner, and T. Grüning, "Rescoring Sequence-to-Sequence Models for Text Line Recognition with CTC-Prefixes," *arXiv preprint arXiv:2110.05909*, 2021.

[46] M. Li, T. Lv, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "TrOCR: Transformer-based optical character recognition with pre-trained models," *arXiv preprint arXiv:2109.10282*, 2021.

[47] U.-V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39–46, 2002.

[48] E. Augustin, M. Carré, E. Grosicki, J.-M. Brodin, E. Geoffrois, and F. Prêteux, "RIMES evaluation campaign for handwritten mail processing," in *Proceedings of the International Workshop on Frontiers in Handwriting Recognition*, 2006.

[49] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character hmms," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 934–942, 2012.

[50] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription alignment of Latin manuscripts using hidden Markov models," in *Proceedings of the International Workshop on Historical Document Imaging and Processing*, 2011.

[51] V. Romero, A. Fornés, N. Serrano, J. A. Sánchez, A. H. Toselli, V. Frinken, E. Vidal, and J. Lladós, "The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition," *Pattern Recognition*, vol. 46, no. 6, pp. 1658–1669, 2013.

[52] J. A. Sánchez, V. Romero, A. H. Toselli, and E. Vidal, "ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (HTRtS)," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 2014.

[53] J. A. Sanchez, V. Romero, A. H. Toselli, and E. Vidal, "ICFHR2016 competition on handwritten text recognition on the READ dataset," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 2016.

[54] T. Strauß, G. Leifert, R. Labahn, T. Hodel, and G. Mühlberger, "ICFHR2018 competition on automated text recognition on a READ dataset," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 2018.

[55] N. Serrano, F. Castro, and A. Juan, "The RODRIGO Database," in *Proceedings of the Language Resources and Evaluation Conference*, 2010.

[56] D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. Ramos Terrades, and A. Juan, "The GERMANA Database," in *Proceedings of the Language Resources and Evaluation Conference*, 2010.

[57] S. Johansson, G. N. Leech, and H. Goodluck, *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computer*. Department of English, University of Oslo, 1978.

[58] T. Causer and V. Wallace, "Building a volunteer community: results and findings from transcribe bentham," *Digital Humanities Quarterly*, vol. 6, 2012.

[59] A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stolz, "Automatic transcription of handwritten medieval documents," in *Proceedings of the International Conference on Virtual Systems and Multimedia*, 2009.

[60] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[61] J. Puigcerver and C. Mocholí, "PyLaia," https://github.com/jpuigcerver/PyLaia, 2018.

[62] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the International Conference on Learning Representations*, 2021.

[63] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the International Conference on Machine Learning*, 2021.