applied sciences

MDPI

Article

Optimal Training Dataset Preparation for AI-Supported Multilanguage Real-Time OCRs Using Visual Methods

Attila Biró [1,2,3,*], Sándor Miklós Szilágyi [1] and László Szilágyi [4,5]

[1] Department of Electrical Engineering and Information Technology, George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures, Str. Nicolae Iorga, Nr. 1, 540088 Targu Mures, Romania; sandor.szilagyi@umfst.ro
[2] Department of Physiotherapy, University of Malaga, 29071 Malaga, Spain
[3] Biomedical Research Institute of Malaga (IBIMA), 29590 Malaga, Spain
[4] Physiological Controls Research Center, Óbuda University, Bécsi út 96/B, 1034 Budapest, Hungary; szilagyi.laszlo@uni-obuda.hu
[5] Computational Intelligence Research Group, Sapientia Hungarian University of Transylvania, 540485 Targu Mures, Romania
* Correspondence: abiro@uma.es; Tel.: +36-30-999-0222

Abstract: In the realm of multilingual, AI-powered, real-time optical character recognition systems, this research explores the creation of an optimal, vocabulary-based training dataset. This comprehensive endeavor seeks to encompass a range of criteria: comprehensive language representation, high-quality and diverse data, balanced datasets, contextual understanding, domain-specific adaptation, robustness and noise tolerance, and scalability and extensibility. The approach aims to leverage techniques like convolutional neural networks, recurrent neural networks, convolutional recurrent neural networks, and single visual models for scene text recognition. While focusing on English, Hungarian, and Japanese as representative languages, the proposed methodology can be extended to any existing or even synthesized languages. The development of accurate, efficient, and versatile OCR systems is at the core of this research, offering societal benefits by bridging global communication gaps, ensuring reliability in diverse environments, and demonstrating the adaptability of AI to evolving needs. This work not only mirrors the state of the art in the field but also paves new paths for future innovation, accentuating the importance of sustained research in advancing AI's potential to shape societal development.

Keywords: multilingual OCR; RNN; CNN; CRNN; SVTR; real-time translation; collaborative diagnostics; real-time text detection; assessment

check for updates

Citation: Biró, A.; Szilágyi, S.M.; Szilágyi, L. Optimal Training Dataset Preparation for AI-Supported Multilanguage Real-Time OCRs Using Visual Methods. *Appl. Sci.* **2023**, *13*, 13107. https://doi.org/10.3390/app132413107

Received: 26 October 2023
Revised: 27 November 2023
Accepted: 5 December 2023
Published: 8 December 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The area of remote diagnosis has undergone a transformational change thanks to the power of cutting-edge collaboration tools, ushering in a new era of multilingual cooperative working sessions that span many sectors, most notably health care [1]. The quest for the establishment of non-invasive, secure, and remote cooperation models is gaining momentum, the apex of which is to find a solution that facilitates real-time sharing of documents and presentations, thereby amplifying the efficacy of these cooperative models [2].

Previously, the application of optical character recognition (OCR) solutions posed a substantial challenge when mixcode [3,4], multiple or synthesized languages (a mixture of context-sensitive languages and Anglo-Saxon and Latin language families), dialects, or varying degrees of technical linguistic proficiency were involved in the process [5]. Furthermore, the risks associated with misinterpretations induced by incorrect translations and deficiencies in the domain knowledge of interpreters cannot be understated.

This paper outlines and tackles five research questions (RQs), which collectively guided the investigations into the optimal preparation of training datasets for AI-supported

multilanguage real-time OCRs in collaborative on-screen solutions. The RQs address key challenges in the OCR technology research domain, acknowledging the complexities introduced by synthesized languages, contextual dependencies, and linguistic diversity within the Latin and Anglo-Saxon language families. The RQs with their specific rationales and contributions are enumerated below:

1. RQ 1: What are the most effective strategies for creating a synthetic training dataset that accurately captures the intricacies of multilanguage OCR, particularly when dealing with context-sensitive languages alongside Latin and Anglo-Saxon language families? Rationale: This RQ addresses the fundamental challenge of preparing an optimal training dataset that accommodates a diverse set of languages, including those with varying contextual dependencies. Based on our research experiments conducted using a convolutional recurrent neural network (CRNN) and scene text recognition with a single visual model for scene text recognition (SVTR) models with PaddleOCR, the need for specific strategies to capture nuances in languages such as Chinese or Japanese (context-sensitive) and English/Hungarian (Anglo-Saxon/Latin) was recognized within a single OCR system for collaborative on-screen solutions.
2. RQ 2: How can the variability of fonts, writing styles, and document layouts be systematically incorporated into the generation of synthetic training data to ensure the robustness of a multilanguage OCR system? Rationale: This question delves into the critical issue of accommodating diversity in fonts, writing styles, and document layouts within a synthetic dataset. Addressing this challenge is crucial for training an OCR system that can effectively handle visual heterogeneity across languages and document types.
3. RQ 3: What are the optimal techniques for metadata-based quality control in the context of a multilanguage OCR training dataset, and how can these techniques be applied efficiently for languages with distinct linguistic features? Rationale: This question focuses on the often overlooked aspect of the advanced level of quality control in multilanguage OCR training datasets. It emphasizes the need for tailored techniques that consider the linguistic peculiarities of each language, ensuring high-quality annotation methods across diverse linguistic backgrounds.
4. RQ 4: What methods and standards can be set up to thoroughly test the precision, dependability, and effectiveness of AI-powered real-time multilanguage OCR systems, taking into account the complexity of synthesized languages and different language families? Rationale: This question addresses the crucial aspect of evaluation in the context of multilanguage OCR systems. It stresses the need for standardized testing methods and standards that take into account the difficulties that come with mixed languages and different types of language so that the performance evaluations of OCR systems can be trusted and compared.
5. RQ 5: What are the effective adaptation and fine-tuning techniques for deep learning models, specifically SVTR and CRNNs, to address the challenges of multilanguage OCR? Furthermore, what strategies can be employed to optimize the performance of these models for real-time on-screen OCR in collaborative environments? The purpose of this inquiryis to examine the process of adapting and optimizing deep learning models, namely SVTR and CRNN, for the purpose of performing multilanguage OCR jobs. The statement recognizes the necessity of adapting models to suit the complexities of various languages and the demands of real-time, collaborative on-screen OCR applications.

We propose Hypothesis 1 (H1) about RQ1: Adding language-specific context modeling techniques to a synthetic training dataset will have a considerable effect on the multilanguage OCR system's accuracy and ability to adapt to different contexts. This is especially true for languages like Japanese and Chinese, which have a lot of complex contextual dependencies, as well as the Latin and Anglo-Saxon language families. It is suggested in H1 that adding language-specific contextual information to the creation of synthetic training data will improve the accuracy of optical character recognition (OCR) for languages

like Chinese and Japanese, which are sensitive to context. Additionally, this approach is expected to maintain a high level of performance for languages belonging to the Latin or Anglo-Saxon language families. The assumption is made that employing a context-aware strategy will effectively reduce the influence of linguistic variations, thereby leading to improved language recognition capabilities across different languages.

As a result of RQ2, Hypothesis 2 (H2) is true: Using a variety of fonts, writing styles, and document layouts when creating fake training data will make the multilanguage OCR system more resilient and adaptable, allowing it to handle the visual differences seen in different languages and document formats. Adding different fonts, writing styles, and document layouts to fake datasets will help an optical character recognition (OCR) system better handle the different visual features found in multilingual and multimodal documents. With this method, it is postulated that the OCR system's capacity for effective generalization can be enhanced through exposure to a wide range of visual features during training.

Similarly, Hypothesis 3 (H3) is deduced from RQ3: The utilization of language-aware techniques for the generation of labels and quality control during the creation of multi-language OCR training datasets will yield enhanced accuracy and linguistic precision in annotations. Consequently, this will lead to improved OCR performance in languages that possess distinct linguistic characteristics. The argument for H3 comes from the idea that using language-specific methods to make labels and make sure that they are accurate will result in training datasets that accurately reflect the unique features of each language. The assumption, based on the experiments, is that the utilization of exact linguistic annotations will result in improved accuracy in recognizing languages that possess different linguistic characteristics.

The fourth idea, Hypothesis 4 (H4) for RQ4, suggests that it will be easier to test AI-powered multilanguage real-time OCR systems in a way that is both rigorous and meaningful if there are standardized testing methods and standards that take into account the complexity of synthesized languages and how different languages work. This, in turn, will enhance their reliability and comparability. The motivation for H5 revolves around the proposition that the credibility of OCR system assessments can be improved by creating evaluation procedures and benchmarks specifically designed to address the unique challenges posed by synthetic languages and different linguistic families. It is thought that using standard testing methods will provide useful information about how well multilingual OCR systems work, which will help people make smart choices about their development and use.

Finally, Hypothesis 5 (H5) is based on RQ5: We can make deep learning models, such as SVTR and CRNNs, better at multilanguage OCR tasks by changing and customizing them to handle the unique language and image needs of those tasks. We expect that this will lead to models that are optimized and able to achieve high accuracy in real-time, collaborative on-screen OCR. The reasoning behind H5 is based on the idea that improving deep learning models like SVTR and CRNNs (used in our experiments) addresses the problems that come with multilanguage or synthesized OCRs, like different kinds of languages or made-up languages, making the models work better. The customization of the models facilitated the efficient management of linguistic and visual intricacies in real-time, collaborative on-screen optical character recognition (OCR) situations.

Presently, the unprecedented advances in artificial intelligence (AI) with statistical ML [6] present an exceptional opportunity for society. In the OCR field, real-time AI models [7] offer a potential solution [8] to overcome these long-standing obstacles [9]. By implementing a near-real-time processing pipeline, AI enables us to confront and triumph over the challenges inherent in multilingual collaboration [10] (e.g., multilingual, multiuser conferences; technical meetings between different domains; and remote diagnosis sessions), sparking a revolution in remote, non-invasive cooperation. This method is unique because it uses AI to help people communicate even though they do not speak the same language. This shows how cutting-edge technology can be used for good, which changes the way we work together globally [11].

1.1. Steps of Training Dataset Generation

Generating synthetic training data for a multilanguage OCR system involves creating synthetic images that simulate the real-world conditions under which the OCR system is expected to operate. This can be done by manipulating various parameters of the image, such as the font type, size, color, background texture, and noise levels. The generated images should be representative of the various languages and writing systems that the OCR system is expected to recognize. Here are some methods and mathematical techniques commonly used for generating synthetic training datasets for a multilanguage OCR system:

1. Text Generation Process: The first step is to generate the text that will be used in the synthetic images. This can be done by using natural language processing techniques such as language models or by collecting and curating text datasets for each language.
2. Font Generation Process: To generate realistic synthetic images, it is important to use fonts that are representative of the languages and writing systems. There are several methods for creating fonts, including the use of generative adversarial networks (GANs) [12,13] or variational autoencoders (VAEs) [14] or simply choosing fonts from libraries that already exist.
3. Image Generation Process: Once the text and fonts have been generated, the synthetic images can be created. This can be done using image manipulation techniques such as rotation, scaling, and translation and by adding various types of noise, such as blur, salt-and-pepper noise [15–18], or Gaussian noise.
4. Augmentation Process: The generated images can be enhanced by using a variety of ways to further diversify the training dataset. This includes techniques such as random cropping, flipping, and color jittering.
5. Labeling Process: Finally, the synthetic images must be labeled with the correct character or text sequence that they represent. This can be accomplished by either manually labeling the photos or by utilizing optical character recognition techniques to extract the text from the artificial images.

Mathematical techniques such as *probability distributions*, *random number generation*, and *matrix transformations* are used extensively in generating synthetic images with realistic noise and other image characteristics, adjusted specifically in the ways required by specific use cases. For example, a Gaussian distribution can be used to add random Gaussian noise to an image, while affine transformations can be used to rotate, scale, or translate an image.

In summary, generating synthetic training datasets for a multilanguage OCR system involves a combination of text and font generation, image manipulation, image augmentation, and labeling techniques. Various mathematical techniques are used to manipulate image characteristics and generate realistic noise, which helps train the OCR system to recognize characters and text sequences in real-world scenarios.

1.2. Alternative Approaches and Differences

A vectorizer in the context of natural language processing (NLP) is a component that transforms textual data into numerical vectors, making them suitable for machine learning algorithms. Vectorization techniques include bag of words (BoW), which represents each document as a vector of word frequencies, and TF-IDF, which represents the importance of a term in a document relative to a corpus. Vectorizers capture the lexical information of the text but may not capture semantic relationships between words. Word embedding is a more advanced vectorization technique that captures semantic relationships between words by mapping words to dense, continuously valued vectors in a high-dimensional space. Techniques like Word2Vec, GloVe, and fastText are used to create word embeddings. Word embeddings enable the representation of words with similar meanings as vectors that are close to each other in the vector space. They have been instrumental in improving the performance of various NLP tasks. Term frequency–inverse document frequency (TF-IDF) is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus). It combines two components: term frequency (TF),

which measures how often a term occurs in a document, and inverse document frequency (IDF), which measures how unique or rare a term is across the corpus. TF-IDF assigns higher weights to terms that are frequent in a document but rare in the corpus, thereby emphasizing the importance of discriminative terms. Vectorization methods such as BoW, TF-IDF, and word embeddings are important in NLP and text analysis. However, the method for preparing the best training dataset for AI-supported multilingual real-time OCRs using visual methods is very different in a number of important ways:

1. *Focus on OCR training data preparation*: Our approach primarily focuses on the preparation of synthetic training datasets for OCR tasks. It addresses the challenges associated with creating datasets that encompass multiple languages, including context-sensitive languages and those from the Latin and Anglo-Saxon language families. The primary goal is to optimize OCR system training, ensuring high accuracy across diverse languages and visual contexts.
2. *Visual methods*: The presented method emphasizes the incorporation of visual elements such as fonts, writing styles, and document layouts into the synthetic dataset generation process. This visual aspect is particularly relevant for OCR, where the recognition of characters and text from images is critical. Traditional NLP techniques like word embeddings and TF-IDF do not directly address these visual considerations.
3. *Multimodal data considerations*: OCR involves the analysis of both textual and visual information. The suggested method takes into account both linguistic and visual features in the training dataset. This lets OCR systems handle documents that are written in more than one language or format.
4. *Language-specific challenges*: Our method recognizes the linguistic and contextual differences among languages, especially when dealing with context-sensitive languages and synthesized languages. It aims to tailor the dataset preparation process to address these language-specific challenges, ensuring that the OCR system performs optimally across diverse linguistic backgrounds.
5. *Application in near-real-time collaborative on-screen OCR* : The proposed approach targets real-time OCR applications within collaborative on-screen solutions. This context imposes stringent requirements on processing speed and accuracy, and the approach is designed to meet these specific demands.

In summary, while vectorization techniques and methods like word embeddings and TF-IDF are fundamental in NLP, our approach stands apart by addressing the intricate challenges posed by multilanguage OCR, emphasizing visual elements, accommodating diverse linguistic and contextual variations, and targeting real-time collaborative OCR applications.

1.3. Convergence Method of Optimal Dataset Preparation

The analysis of model results from experiments involves the utilization of a specialized pivot table based on metadata. This pivot table provides insights into the error ratio associated with different types of feature sets, thereby indicating the relative significance of these sets. Additionally, it identifies areas where additional data injection is required to improve the model's accuracy. The pivot encompasses several key aspects, namely "Formatting", "Fonts", "Word count", and "Bad recognitions", while also allowing for the option of manual analysis. In the context of metadata-based evaluation, the process involves loading the predictions into a Jupyter notebook. Subsequently, the notebook applies a filtering mechanism to identify any missing predictions. Finally, the predictions are merged with the metadata table. The pre-existing pivot table is populated with the data, resulting in an updated table incorporating the most recent information. This includes the source data, along with the accompanying metadata and forecast. The model and its corresponding epoch are appended to the table to ensure comprehensive representation. The chosen model's PaddleOCR and Jaccard scores were calculated and plotted in a diagram. Jaccard scoring was also used to evaluate the different text formatting options by category, including alignment, background color, text color, blur, noise, italic, bold, etc. If the Jaccard score is deemed reliable, the model performs beyond 85%. Similarly, when the

actual field is indicated as green, the model achieves a Jaccard score surpassing 85% in a significant proportion of instances. Additionally, the accuracies of typefaces, languages, word length, and character length, which may be sorted by language, are graphically represented. If the Jaccard score is deemed valid, the model demonstrates performance over 85%. Similarly, if the valid field is highlighted in green, the model achieves a Jaccard score above 85% in a significant proportion of instances.

1.4. Rationale behind the New Approach

Various multilanguage OCR systems are available in the market already with publicly accessible multi-language databases. These systems still focus on classic OCR use cases, without providing a reliable solution for on-screen, digital OCR to support real-time, multilanguage, or cross-sectoral collaborative, supportive works, workshops, and multi-geolocation issues. In our research, we chose and expanded one of the most reliable OCR solutions: PaddleOCR (Figure 1), a practical ultra-lightweight OCR system supporting more than 80 languages.

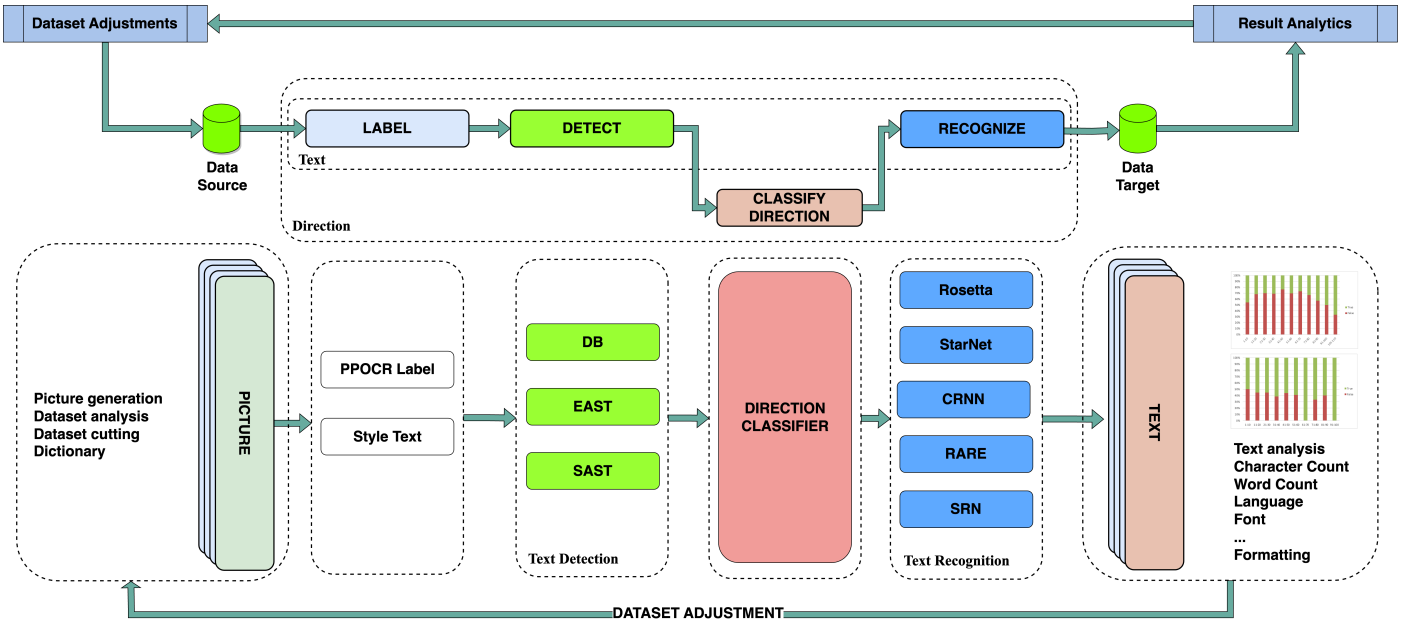

Figure 1. Adjusted PaddleOCR architecture (adapted from [19]).

PaddleOCR [20] provides promising results for Far-East languages (like Japanese and Chinese) and handles the multilanguage approach with flexible, adaptable AI models [21]. Based on our research, not even PaddleOCR was able to manage accurate phrase recognition in multilingual systems or for synthetic languages used in the IT or sports sector for on-screen simulations.

The research analyzing the existing solutions available on the market suggests that there is currently no magic formula that can be applied to prepare an optimal training dataset that covers all of the on-screen collaboration tool requirements. Every use case is different, so it is suggested to handle the topics as standalone projects. The training dataset volume will be defined by the use case, the involved languages, and the specific vocabulary to be used. Existing annotated datasets are not golden sets, but they can be used when preparing the method that covers all aspects of our requested OCR system [21].

The formulation of an optimal, vocabulary-based training dataset for multilingual, real-time OCR systems necessitates a strategic approach to various key elements. These include the careful selection of adaptive vocabulary, implementation of cross-lingual pretraining [22], deployment of dynamic dataset augmentation strategies, ensuring domain-specific adaptation, addressing data imbalance issues, applying multimodal learning, and establishing appropriate evaluation metrics and benchmarks. [23]. A data scientist who addresses

these novel aspects can build OCR systems that are accurate and efficient and capable of handling a wide range of languages and contexts.

1.5. Proposed Improvements

To prepare an optimal, vocabulary-based training dataset distribution for multilingual OCR training, there are several aspects and approaches that need to be taken into consideration, as follows:

1. *Dynamic Dataset Augmentation* [24,25]: The training dataset should be continually updated and expanded to include new text samples (using the Sobel filter [26]), variations in language, and emerging writing styles. This can be achieved through techniques such as synthetic data generation [27], web scraping, and crowd sourcing. Dynamic dataset augmentation ensures that the OCR system remains accurate and up-to-date with evolving language trends.
2. *Adaptive Vocabulary Selection* [28]: To cater to the diverse nature of languages and scripts, an adaptive vocabulary selection method is crucial. This approach enables the identification and selection of the most relevant and frequently occurring words, characters, or n-grams in each language, providing an optimal basis for training the OCR system.
3. *Domain-specific Adaptation* [29]: To enhance the OCR system's performance in specific industries or use cases, the training dataset should be enriched with domain-specific data. For example, incorporating legal, medical, or technical jargon can help improve the system's performance in those respective fields.
4. *Cross-lingual Pretraining* [30]: Leveraging transfer learning from a pretrained multilingual model [31] can aid in raising OCR system's efficiency for low-resource languages. By training on a large, diverse, and high-quality dataset, the OCR model learns to identify common patterns and structures across different languages, leading to better generalization.
5. *Data Imbalance Mitigation* [32]: A balanced dataset (Figure 2) is essential in training a robust multilingual OCR system. This involves addressing class imbalances by oversampling underrepresented classes or undersampling over-represented classes. Additionally, incorporating data augmentation techniques [33] such as rotation, scaling, and noise addition can help increase the variability of the training data.
   In the case of multilanguage training involving Japanese, English, and Hungarian at the same time, the character occurrence frequency must be evaluated per domain and per targeted solution.
6. *Evaluation Metrics and Benchmarks* [34]: Establishing appropriate evaluation metrics and benchmarks for multilingual OCR systems [35] is vital for assessing performance, identifying areas for improvement, and facilitating model comparison. These metrics should include character- and word-level recognition rates, language identification accuracy, and domain-specific evaluation measures.
7. *Multimodal Learning* [36]: Incorporating additional contextual information such as domain-specific metadata can improve the OCR system's ability to recognize text within different contexts. Multimodal learning enables the model to leverage multiple sources of information to enhance its understanding of the text.

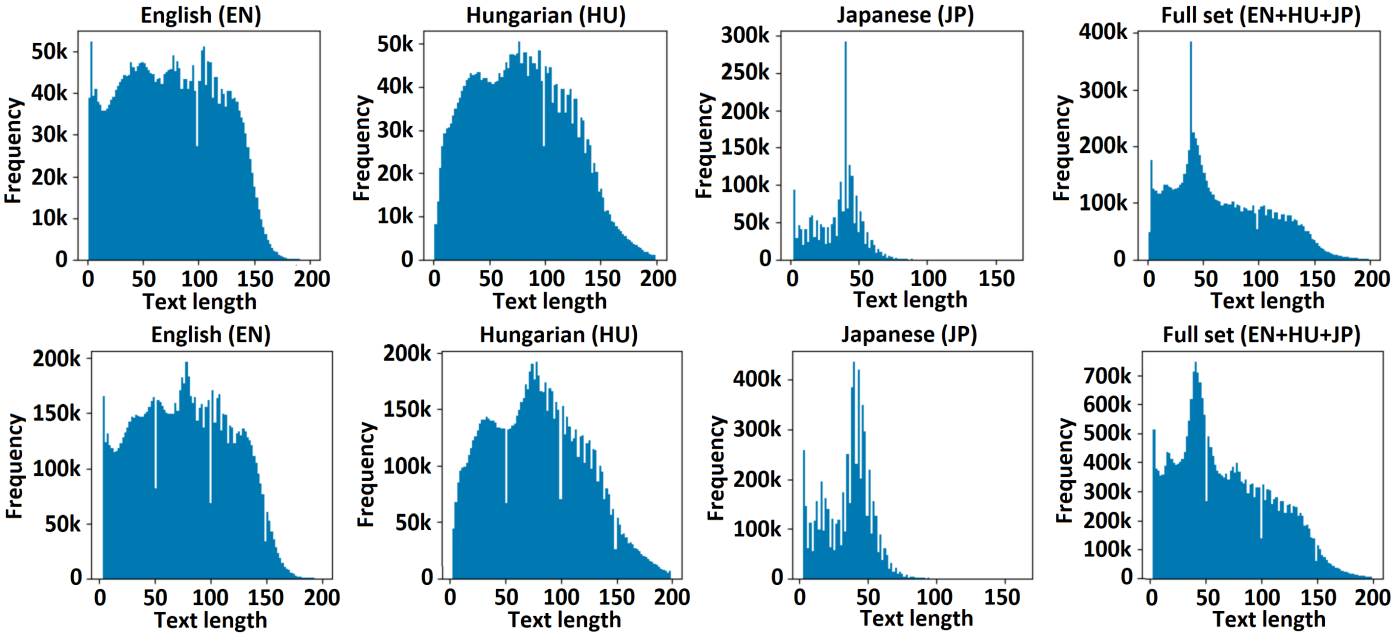

Figure 2. Result of data imbalance mitigation.

2. Objectives

This study introduces a comprehensive methodology that synergistically integrates cutting-edge approaches, culminating in a significant time-efficient tool for OCR developers. By facilitating the construction of optimal, purpose-specific datasets, this approach paves the way for the development of high-precision models. The primary objectives of this research project can be formulated as follows:

1. To develop an optimal vocabulary-based training dataset for multilingual, AI-powered, real-time OCR systems, fostering technical excellence and pushing the boundaries of the research field. Several modern text recognizers are presented in Figure 3, all of which not only have high accuracy but also represent fast inference speeds: (a) CNN-RNN-based models [37]; (b) encoder–decoder models [38] involving multihead self-attention (MHSA) [39] and multihead attention (MHA) mechanisms [40]; (c) vision-language models [41]; and (d) SVTR [42], which recognizes scene text by using a single visual model built with cross-lingual capability [7] in mind.
2. To ensure comprehensive language representation in existing (e.g., PaddleOCR) and future custom OCR systems, with an initial focus on English, Hungarian, and Japanese as representative languages but with the flexibility to extend the methodology to any existing or synthesized languages.
3. To secure and control the requirements of high-quality and diverse data samples, maintaining a balanced dataset that can effectively support the development of accurate, efficient, and versatile OCR systems.
4. To enhance contextual understanding and facilitate domain-specific adaptation in OCR systems (e.g., health care, IT, and industrial or cross-domain topics), enabling them to adapt to a myriad of real-world scenarios.
5. To promote robustness and noise tolerance in OCR systems, enhancing their reliability and performance in diverse environmental conditions.
6. To enable scalability and extensibility in OCR systems, demonstrating the potential of AI to grow and adapt with evolving societal needs (e.g., real-time, multilingual, cross-domain interpreter systems).
7. Through the use of methods like CNN, RNN, CRNN, and SVTR, it is possible to find a visually controlled solution that streamlines the creation of training datasets and vocabulary for real-time OCR projects (such as PaddleOCR).

2.1. Optimal Training Dataset Preparation

It is always challenging to prepare an optimal training dataset because it is subject to the following combined requirements: (1) comprehensive language representation [43], (2) high-quality data preparation [44], (3) diverse data sample utilization [45], (4) balanced dataset usage [46]—especially in the case of multi-language training [47,48], (5) contextual understanding, (6) domain-specific adaptation, (7) scalability and extensibility, and (8) robustness and noise tolerance.

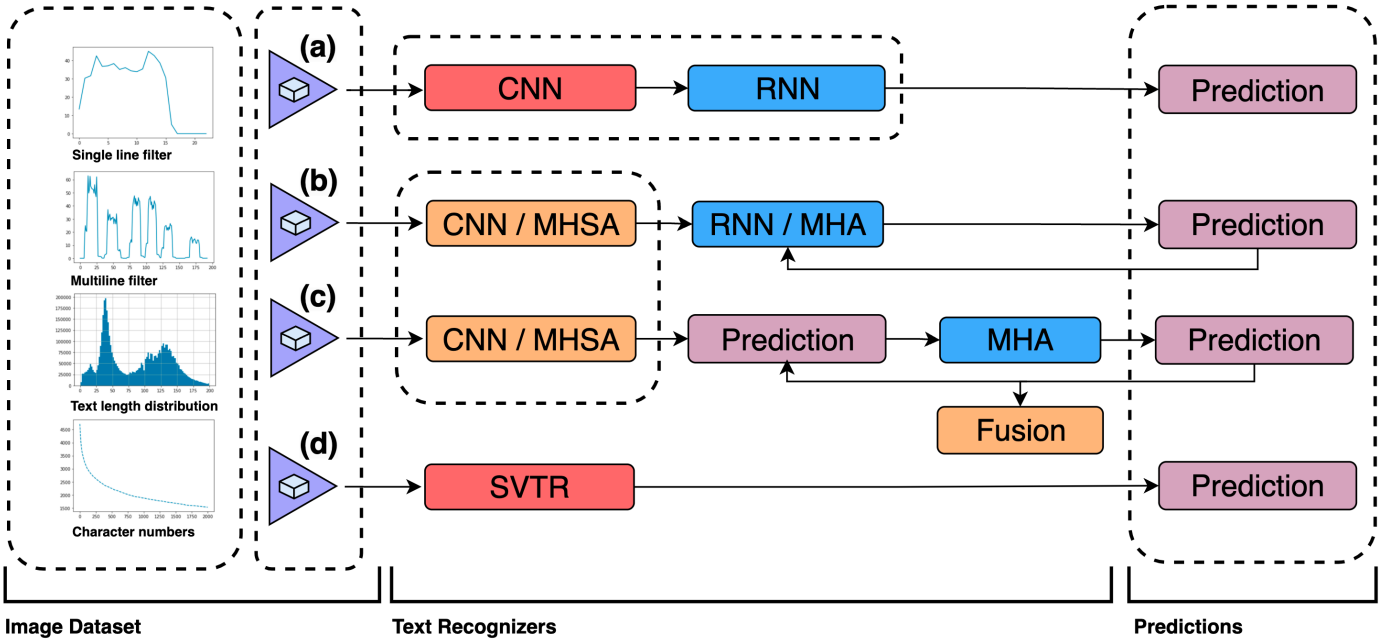

Figure 3. Dataset preparation for real-time OCR (adapted from [42,47]).

The training data reported in this paper were generated with a custom data generator tool developed in Python [1], which can manage customized templates to obtain real-life-ready annotated images with metadata for further analysis. The tool can specify and/or customize the precise text sources to use and the proportions of these text sources to appear in the produced training data [1]. For multilanguage training, the proper text distribution and optimal vocabulary preparation are two of the most important areas where we saw room for improvement in the case of the existing approaches.

2.2. Novelty of Distribution Analysis and Vocabulary Management

The proposed visual controller method, encompassing vocabulary, distribution, and single-/multiline dataset generation facilitates the precision tuning of the multilingual vocabulary set and distribution preparation. This precision, in turn, enhances recognition accuracy across a wide range of languages and scripts. The OCR algorithm is able to recognize and assimilate certain visual aspects of each language by skillfully selecting different and representative examples for the training dataset. This leads to a marked increase in text recognition accuracy and a decrease in errors, setting it apart from conventional methods.

The novelty lies in the system's adaptive capability. Through the continual monitoring of the performance of the OCR model and the refinement of the multilingual vocabulary set and distribution preparation, the system demonstrates a persistent enhancement in performance. Moreover, it demonstrates remarkable flexibility to accommodate new languages, scripts, or domain-specific text, thereby embodying a significant leap forward in OCR technology.

3. Materials and Methods

To improve the performance of AI-supported OCRs, it is essential to provide diverse and representative training data. One way to achieve this is through *multimodal data augmentation techniques*, which involve generating new training samples by combining text, images, and other media. This can include simulating various fonts, sizes, and orientations of text; adding noise and artifacts to mimic real-world conditions; and incorporating context from images, such as background elements or text within images, to improve OCR accuracy. *Traditional training datasets* are often static and may not evolve with the changing requirements of the OCR system.

3.1. Single-Line/Multiline Regularization

It is important to handle single- and multiline pictures properly; this is the first point where it is possible to apply a visual control to the data generation process. To optimize the dataset, we used the Sobel filter [49] on top of the generated dataset to identify, extract, and annotate the multiline pictures for further analysis. Figure 4 showcases the application of the Sobel filter, wherein the horizontal axis represents the rows of the image, and the vertical axis represents the number of bright pixels. When using this filter, the algorithms can determine whether the image contains multiple rows based on the presence of peaks on the chart if a large number of paler pixels is present in a given pixel row. Consequently, it includes characters. See the statistics of a single-line image and the statistics of a multiline image with six text lines below.

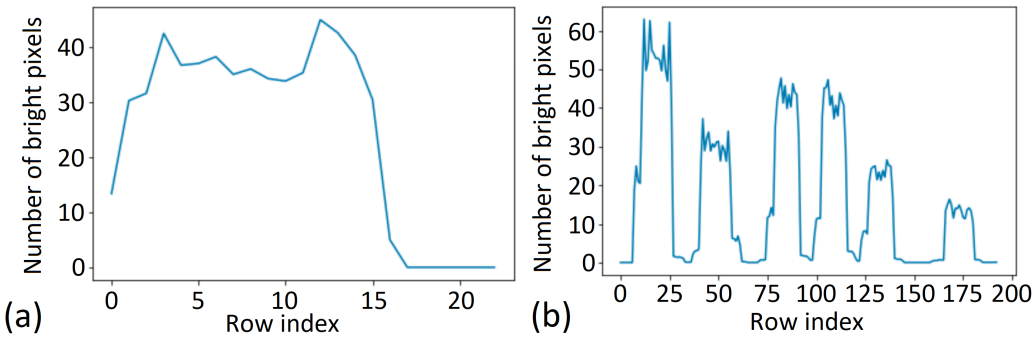

Figure 4. Sobel filter on data generation: (**a**) in the case of single-line text; (**b**) in the case of multiline text.

Active learning is a technique that helps to identify areas where the model is uncertain and prioritizes those areas for human annotation. By incorporating a human-in-the-loop approach [50], we can continuously refine the training dataset based on model performance, focusing on the most difficult or ambiguous cases. This enables the OCR system to better adapt to real-time requirements and ultimately leads to a more accurate and robust model.

Multilanguage, real-time OCR systems in domain-specific use cases (e.g., collaborative work or remote diagnosis [1]) need to recognize and understand texts in various languages, often with limited or imbalanced training data for each language [47]. Cross-lingual and cross-domain transfer learning techniques [51] can be used to transfer the understanding of one language or field to another, improving the model's performance on low-resource languages or underrepresented domains. By pretraining the model on large, diverse corpora of multilingual text, it can be fine-tuned to specific languages or domains with limited data, leading to a more efficient and accurate OCR system.

3.2. Optimal Data Distribution Calculation

3.2.1. Data Preprocessing

The prerequisites for OCR training include preprocessing the data and ensuring their cleanliness and correct formatting. The following formula represents the data preprocessing step:

$$x_{pre}^{(i)} = \text{Preprocess}(x^{(i)}), \tag{1}$$

where $x^{(i)}$ represents the ith training sample, and $x_{pre}^{(i)}$ represents the preprocessed version of the sample.

3.2.2. Feature Extraction

Ref. [52] after preprocessing the data, the next stage is to extract the features that can be utilized to train the OCR system [52]. The following formula represents the feature extraction step:

$$f(x_{pre}^{(i)}) = \text{ExtractFeatures}(x_{pre}^{(i)}), \tag{2}$$

where $f(x_{pre}^{(i)})$ represents the extracted features for the ith training sample.

3.2.3. Model Training

Once the features have been extracted, the OCR system can be trained by using a suitable machine learning algorithm. The following formula represents the model training step:

$$\hat{y}^{(i)} = \text{Train}(f(x_{pre}^{(i)}), y^{(i)}), \tag{3}$$

where $\hat{y}^{(i)}$ represents the predicted output for the ith training sample, $y^{(i)}$ represents the true label for the sample, and Train represents the training algorithm.

3.2.4. Data Distribution Calculation

The following formula can be used to calculate the optimal data distribution for the training of an OCR system for English, Hungarian, and Japanese languages:

$$p_i = \frac{n_i}{\sum_{j=1}^{m} n_j}, \tag{4}$$

where p_i represents the probability of selecting a training sample from language i, n_i represents the number of training samples in language i, and m represents the total number of languages in the dataset [53,54].

By using these formulas in combination with suitable machine learning algorithms and training techniques, we can optimize the data distribution for the training of OCR systems for English, Hungarian, and Japanese languages.

3.3. Training Dataset

For the experiments, we used training datasets of 15 M, 30 M, and 50 M generated phrases with a character number generated between 0 to 200 characters in English, Hungarian, and Japanese. For data generation, we employed a realistic template-based picture generation algorithm written in Python [47].

3.4. Exploratory Data Analysis

It is recommended that the data generation process be an iterative process and that the distribution of the training dataset be visualized per template or application category, as represented in [47]. The visual representation of the text length is the most critical in the case of multilanguage OCR, especially when we combine European language families

(e.g., Anglo-Saxon, Latin, and Germanic) with Asian ones (e.g., Japanese and Chinese). The histogram of generated text lengths is presented in Figure 5.

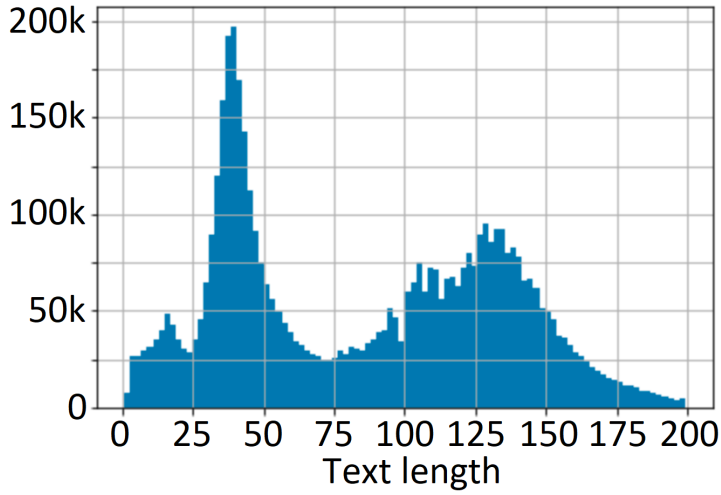

Figure 5. Distribution of text lengths in the train dataset.

3.5. Vocabulary Generation

The method used for both (generated and unified) types of vocabulary generation (see Figure 6) is the one presented in [47]. Experiments suggest that the vocabulary generated from the training dataset provides more accurate results but needs expert supervision of dataset generation. We highly recommend this option because this method ensures that only the characters present in the training dataset will appear in the vocabulary and that there will not be any unnecessary characters at all. This will be an important factor when a Japanese or Chinese character set is included in the OCR development scope.

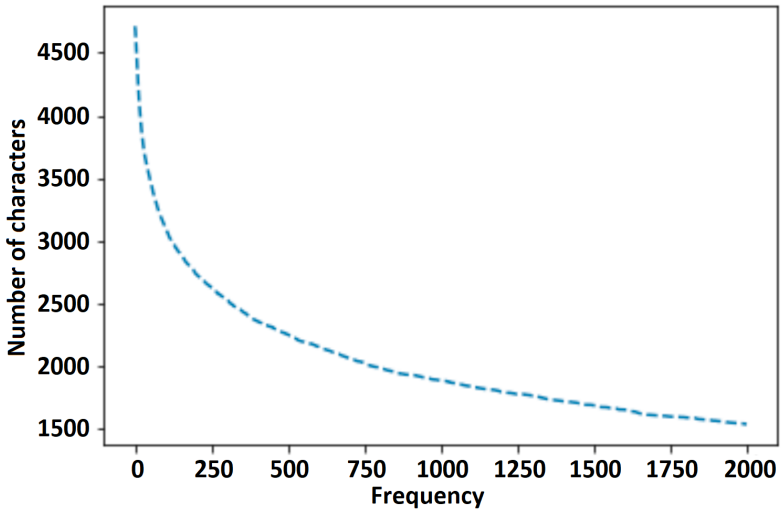

Figure 6. Character number distribution.

4. Experiments

In the conducted trials, both the CRNN and SVTR models were employed. CRNN is a prevalent and extensively used modeling approach in image-based text recognition. Extracting advantageous features involves the consecutive application of convolutions and max-pooling processes. Subsequently, the inputs mentioned above are introduced into the recurrent unit to record and analyze the temporal information effectively. The terminology "v2" (see Tables 1–5) was assigned to these model architectures throughout the trials, as PaddleOCR2 utilized them for text recognition. The SVTR approach initially breaks

down image text into smaller patches called character components. Subsequently, there is a recurring implementation of hierarchical steps, including mixing, merging, and/or combining components at the component level. Global and local mixing blocks were developed to detect intercharacter and intracharacter patterns, resulting in a perception of character components at many levels of granularity. The terminology "v3" (see Tables 1–5) was assigned to these model designs throughout the studies, as PaddleOCR3 specifically employed this model variant for text recognition purposes.

Table 1. Vocabularies used to test the approach.

Name	Vocabulary Name	Size	Type
15M_enhujp_v2_1	training9-200_vocab_min200.txt	2737	Generated
30M_enhujp_v2_4	dashboard_vocab.txt	1441	Generated
50M_enhujp_v2_2	training_min250_nQA.txt	3964	Generated
3M_dashboard_eng_V1	training9-200_vocab_min9500.txt	803	Generated
3M_dashboard_hun_v1	186k_extended_vocab.txt	112	Unified
3M_dashboard_enhujp_v2_1	jpn_latin_dict.txt	4444	Unified
10M_enhujp_v2_1	jpn_latin_dict.txt	4444	Unified
30M_enhujp_v2_1	dashboard_vocab.txt	1441	Generated

Table 2. Vocabularies used for approach validation according to the experiments highlighted in [47].

Project	Name	Vocabulary Name	Size	Vocabulary Type
OCR_hun	155k_hu_v2_2	606k_hun_vocab.txt	112	Generated
OCR_hun	155k_hu_v2_1	606k_hun_vocab.txt	112	Generated
OCR_enhu	SVTR	extended_vocab.txt	201	Unified
OCR_enhu	CRNN	extended_vocab.txt	201	Unified
OCR_multilang	5M_enhujp_v2_8	training9200_vocab_min9500.txt	803	Generated
OCR_multilang	5M_enhujp_v2_6	training9200_vocab_min200.txt0	2737	Generated
OCR_multilang	5M_enhujp_v2_5	training9200_vocab_min9500.txt	803	Generated
OCR_multilang	5M_enhujp_v2_4	training9200_vocab_min200.txt	2737	Generated
OCR_multilang	5M_enhujp_v2_3	jpn_latin_dict.txt	4444	Unified
OCR_multilang	5M_enhujp_v2_2	jpn_latin_dict.txt	4444	Unified
OCR_multilang	5M_enhujp_v2_1	jpn_latin_dict.txt	4444	Unified
OCR_multilang	10M_enhujp_v3_ 1	jpn_latin_dict.txt	4444	Unified
OCR_multilang	5M_enhujp_v3_4	jpn_latin_dict.txt	4444	Unified
OCR_multilang	5M_enhujp_v3_3	jpn_latin_dict.txt	4444	Unified
OCR_multilang	5M_enhujp_v3_2	jpn_latin_dict.txt	4444	Unified
OCR_multilang	4M_enhujp_pre_v3_1	jpn_latin_dict.txt	4444	Unified
OCR_multilang	4M_enhujp_pre_v3_1	jpn_latin_dict.txt	4444	Unified
OCR_multilang	4M_enhujp_v3_5	4M_vocab.txt	4721	Generated
OCR_multilang	4M_enhujp_v3_4	4M_min20_vocab.txt	3980	Generated
OCR_multilang	2M_enhujp_v2_2	2M_min2k_vocab.txt	968	Generated
OCR_hun	186k_hu_v2_1	186k_extended_vocab.txt	112	Unified

Table 3. Data distribution and results.

Name	Steps	Epochs	Best_Acc	Distribution	Vocab Size
15M_enhujp_v2_1	1.32 M	100	85%	Figure 7	2737
30M_enhujp_v2_4	0.7 M	10	98.76%	Figure 8	1441
50M_enhujp_v2_2	1.71 M	10	94%	Figure 9	3964

Table 4. Experimental results of feasibility testing.

Name	Steps	Epochs	Best_Acc	Dataset	Vocab Size
3M_dashboard_eng_V1	1.83 M	90	91%	3 M	803
3M_dashboard_hun_v1	0.96 M	47	89%	3 M	112
3M_dashboard_enhujp_v2_1	1.61 M	100	90%	3 M	4444
10M_enhujp_v2_1	3.96 M	100	85%	8.5 M	4444
15M_enhujp_v2_1	1.32 M	100	85%	15 M	2737
30M_enhujp_v2_4	0.7 M	10	98.76%	30 M	1441
30M_enhujp_v2_1	0.15 M	10	95.87%	30 M	1441
50M_enhujp_v2_2	1.71 M	10	94%	50 M	3964

Table 5. Experimental results and accuracy information on validation testing sets according to the experiments presented in [47].

Project Type	Name	Dictionary	Epochs	Text Length	Training Batch Size	Evaluation Batch Size	Image Shape	Learning Rate	Train Best Acc	Training Size	Evaluation Size
Single language	155k_hu_v2_2	606k_hun_vocab.txt	2000	100	128	128	[3, 32, 512]	0.0050	0.949	77 k	15 k
Single language	155k_hu_v2_1	606k_hun_vocab.txt	2000	100	600	256	[3, 32, 128]	0.0050	0.301	77 k	15 k
Multilanguage	SVTR	extended_vocab.txt	100	150	128	128	[3, 48, 320]	0.0010	0.055	66 k	33 k
Multilanguage	CRNN	extended_vocab.txt	100	250	768	256	[3, 32, 100]	0.0005	0.003	66 k	33 k
Multilanguage	5M_enhujp_v2_8	training9-200_vocab_min9500.txt	100	200	48	48	[3, 32, 1024]	0.0005	0.958	5 M	101 k
Multilanguage	5M_enhujp_v2_6	training9-200_vocab_min200.txt	100	200	64	64	[3, 32, 1024]	0.0050	0.938	5 M	101 k
Multilanguage	5M_enhujp_v2_5	training9-200_vocab_min9500.txt	100	200	48	48	[3, 32, 1024]	0.0050	0.885	5 M	101 k
Multilanguage	5M_enhujp_v2_4	training9-200_vocab_min200.txt	100	200	64	64	[3, 32, 1024]	0.0050	0.922	5 M	101 k
Multilanguage	5M_enhujp_v2_3	jpn_latin_dict.txt	100	200	50	50	[3, 32, 1024]	0.0050	0.780	8.5 M	170 k
Multilanguage	5M_enhujp_v2_2	jpn_latin_dict.txt	100	200	30	30	[3, 32, 2048]	0.0050	0.767	8.5 M	170 k
Multilanguage	5M_enhujp_v2_1	jpn_latin_dict.txt	100	100	50	50	[3, 32, 1024]	0.0050	0.920	5 M	100 k
Multilanguage	10M_enhujp_v3_1	jpn_latin_dict.txt	50	200	60	60	[3, 32, 1024]	0.0010	0.483	8.5 M	170 k
Multilanguage	5M_enhujp_v3_4	jpn_latin_dict.txt	50	100	50	50	[3, 32, 512]	0.0001	0.750	5 M	100 k
Multilanguage	5M_enhujp_v3_3	jpn_latin_dict.txt	50	100	50	50	[3, 32, 512]	0.0001	0.690	5 M	100 k
Multilanguage	5M_enhujp_v3_2	jpn_latin_dict.txt	50	110	50	50	[3, 32, 512]	0.0001	0.300	5 M	100 k
Multilanguage	4M_enhujp_pre_v3_1	jpn_latin_dict.txt	50	200	24	24	[3, 32, 768]	0.0005	0.500	4.4 M	88 k
Multilanguage	4M_enhujp_pre_v3_1	jpn_latin_dict.txt	50	200	28	28	[3, 32, 768]	0.0005	0.429	4.4 M	88 k
Multilanguage	4M_enhujp_v3_5	4M_vocab.txt	50	200	28	28	[3, 32, 768]	0.0010	0.429	4.4 M	88 k
Multilanguage	4M_enhujp_v3_4	4M_min20_vocab.txt	50	200	28	28	[3, 32, 768]	0.0005	0.464	4.4 M	88 k
Multilanguage	2M_enhujp_v2_2	2M_min2k_vocab.txt	100	200	64	64	[3, 32, 1024]	0.0050	0.672	2 M	220 k
Single language	186k_hu_v2_1	186k_extended_vocab.txt	2000	100	128	128	[3, 32, 512]	0.0050	0.961	186 k	34 k

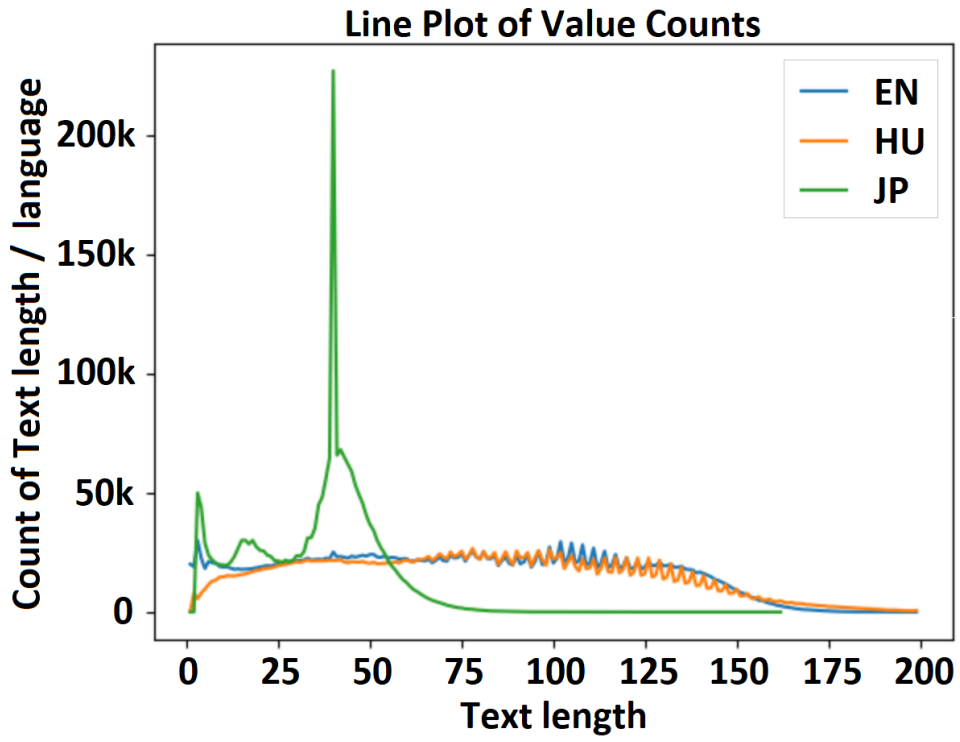

Figure 7. Text length distribution—experiment 1 (15M_enhujp_v2_1): English–Hungarian–Japanese distribution.

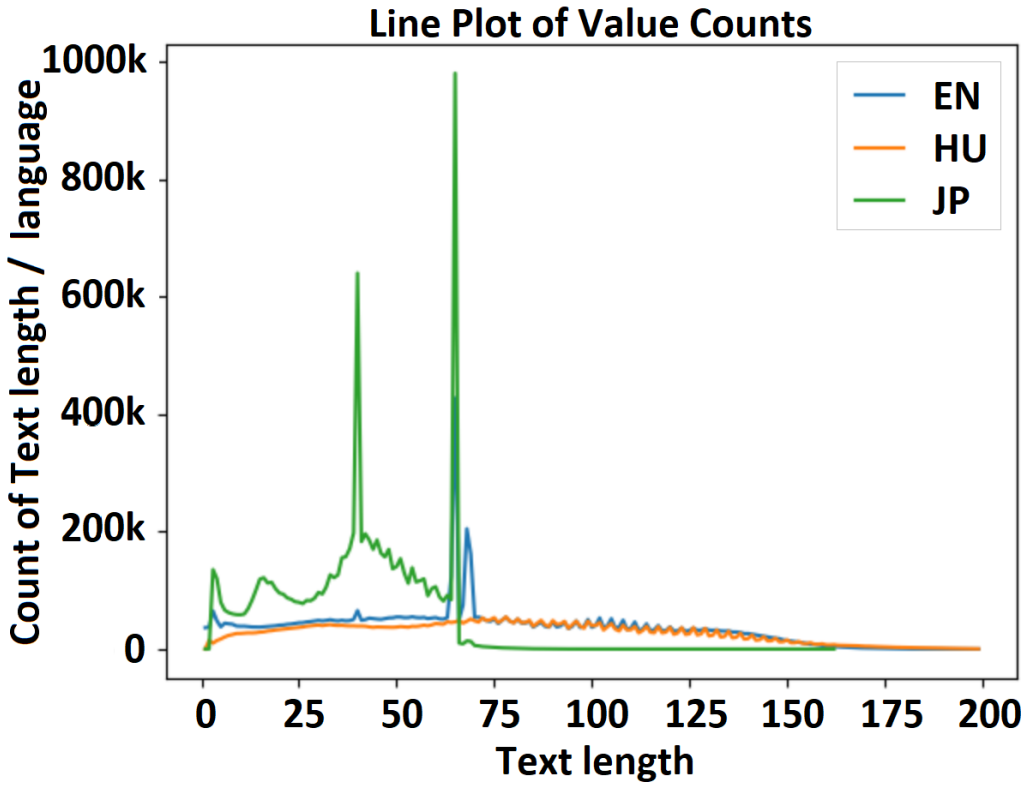

Figure 8. Text length distribution—experiment 1 (30M_enhujp_v2_4): English–Hungarian–Japanese distribution.

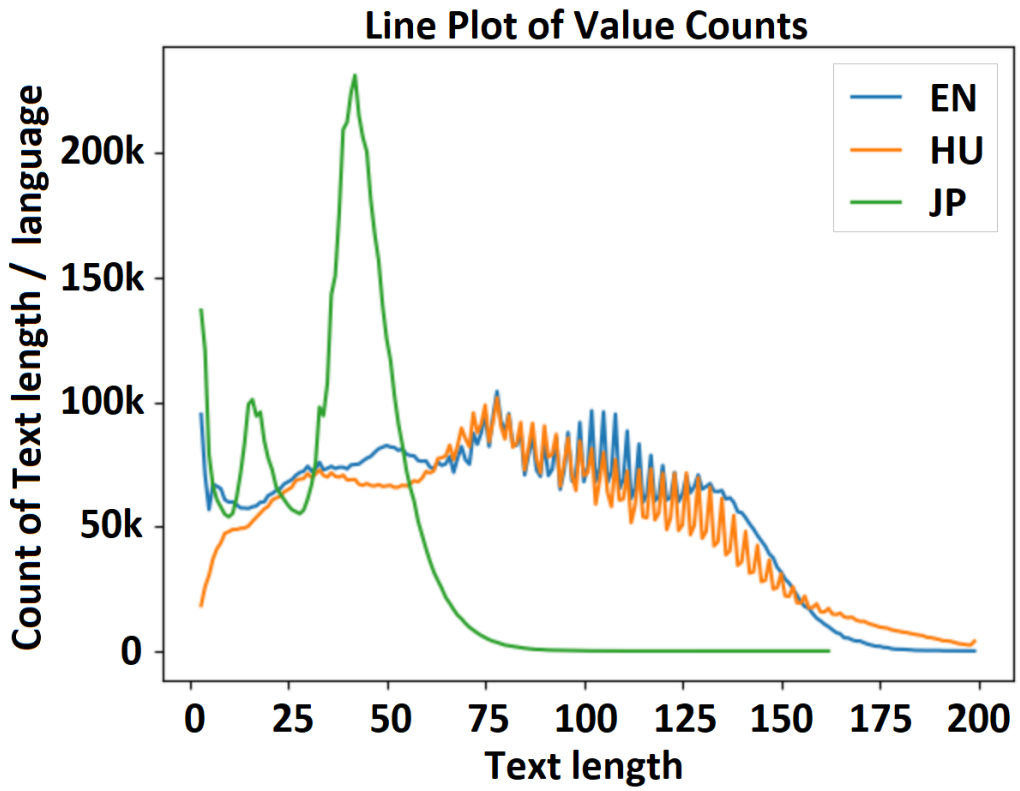

Figure 9. Text length distribution—experiment 1 (50M_enhujp_v2_2): English–Hungarian–Japanese distribution.

We investigated three different project types during the experiments [1], and based on the results, we validated the outcomes with a multitype dataset as presented in [47]. The training category was layered into (1) English; (2) Hungarian; and (3) multilanguage trainings, in which Japanese, English, and Hungarian languages were all present in the dataset. During our experiments, we worked with large datasets as well, but based on the initial result, these experiments were frozen because of the expected running times and are to be continued in high-performance computing (HPC) environment.

The experiments were executed in the following dedicated environments: Config 1 PC with the following configuration: MBO Gigabyte Z390 Aorus Pro, CPU INTEL Core i7-8700K 3.7 GHz 12 MB LGA1151, DDR4 32 GB 3600 MHz Kingston HyperX Predator Black CL17 KIT2, VGA MSI RTX 2080 Ventus 8 GB, SSD M.2 SAMSUNG 970 Pro 1 TB, Corsair RMx (2018) 750 W Modular 80+ Gold) [47]; The second configuration: VGA MSI RTX 2080 Ventus 8 GB (2x); and in the last phase we used a configuration with VGA MSI RTX 2080 Ventus 8 GB (6x).

For the research, we used both generated and unified vocabulary types [47]. Table 1 shows the adaptive vocabulary selection used in the experiments conducted in the feasibility study, while Table 2 represents the validation phase with a wide variety of datasets and dictionaries.

Table 3 shows the details of the "15M_enhujp_v2_1", "30M_enhujp_v2_4", and "50M_enhujp_v2_2" projects. In the experiments, we used the SVTR [42,47] and CRNN [47,55] models.

These figures are accessible to the public; see the Data availability statement. The experiments were conducted with the SVTR [42,47] and CRNN [47,55] models.

Challenges and Mitigation Strategies

Various challenges such as time-consuming procedures, unforeseen issues, and inconsistencies in the observed behaviors characterized the experimental phase. The challenges and constraints faced throughout the execution of PaddleOCR are highlighted below:

1. *Single-line and/or multiline approach*: One of the important technical questions was the capacity of PaddleOCR to handle multiple lines of text. During the experimental phase, it was seen that the system exhibited restricted multiline capabilities under certain conditions. Specifically, this limitation became apparent when the vocabulary size was sufficiently short (ranging from 100 to 200 characters) and the training examples were meticulously chosen. During multilingual experimentation, it was out of scope to provide sufficient time for the manual review of training data that involved character-based multiline analysis. As a result, we opted for the single-line solution offered by PaddleOCR.
2. *Vocabulary generation*: PaddleOCR currently lacks an in-process vocabulary-generating tool. The process of vocabulary formation necessitates manual management. There are two methods by which a lexicon or dictionary can be generated for training. (1) One potential approach is the development of a cohesive and standardized lexicon, commonly referred to as a "golden" vocabulary, which is meticulously overseen and curated by knowledgeable human specialists. (2) An alternative approach involves generating language based on the training data. With this approach, we can guarantee that the vocabulary only includes characters that are found in the training sample, eliminating any extraneous characters.
3. *SVTR vs. CRNN*: During the course of the investigation, PaddleOCR introduced a novel iteration of its software (PaddleOCRv2.5.0 with PP-OCRv3) [56],which incorporates the SVTR paradigm. According to PaddleOCR, the newly proposed model architecture exhibits superior performance compared to its predecessor, CRNN. Numerous experiments were conducted on the novel model architecture, consistently revealing its inferior performance compared to the capabilities exhibited by the prior architecture. Following a thorough analysis of the results obtained using the SVTR, it was determined that reverting back to the utilization of CRNN models is the most

appropriate course of action. In the high-performance computing (HPC) environment, we intend to conduct a comprehensive examination of both methodologies.

4. *Word spacing and segmentation issues*: Japanese texts had improper word spacing during the experiments. Using a tiny-segmenter Japanese text tokenizer improved the performance. Setting up tiny segmenters and tokenizing/segmenting Japanese or Chinese requires expertise. Japanese vs. Unicode character coding is another issue. High-level OCR engines work in a straightforward manner. The model looks at each character to identify it. This may lead to errors. Imagine identical characters with different character codes. In this example, the model will recognize the character as one of two options during training. If this character belongs to another character code, the model returns the information that it made a mistake because it is not that character. This causes training and inference mistakes for the model. The solution is to identify these homoglyphs and change them all to the selected character. For Latin characters, dictionaries solve the problem, but for Japanese or Chinese literature, homoglyphs require further study and depend on the domain. A custom dataset-generating tool handled homoglyphs that way. The 'normalizeHomoglyps' attribute works with a predefined homoglyph table to handle characters/symbols that seem similar but have various character encodings for text sources. The latest version provides automatic Unicode normalization for each text source using the NFKC method.

5. Results

Developing an optimal dataset with well-defined evaluation metrics and benchmarks enables researchers and AI specialists to assess the performance of various OCR models, identify areas for improvement, and facilitate model comparison and selection.

The approach was validated on different datasets, as shown in Table 5, which displays the accuracy of the models, with the most prospective results highlighted in bold. A deeper elaboration of the experiments is presented in [47].

Figure 7 shows a visualization of the distribution of text length in the English, Hungarian, and Japanese datasets for experiment 1 (Experiment1_15M_enhujp_v2_1), where the horizontal axis represents the text length and the vertical axis represents the frequency on each subchart. Here, the mean, median, minimum, and maximum values for the filtered text lengths and for the overall text length without considering the language are as follows. (1) English dataset: mean, 75.404116; median, 75.0, min/max, 1/199. (2) Hungarian dataset: mean, 79.014045; median, 78.0; min/max, 1/99. (3) Japanese dataset: mean, 34.984825; median, 39.0; min/max, 1/162; (4) full (EN-HU-JP) dataset: mean, 65.619248; median, 55.0; min/max, 1/199.

Figure 8 shows a visualization of the distribution of text length in the English, Hungarian, and Japanese datasets for experiment 2 (Experiment1_30M_enhujp_v2_4), where the horizontal axis represents the text length and the vertical axis represents the frequency on each subchart. The mean, median, minimum, and maximum values for the filtered text lengths and for the overall text length without considering the language are as follows. (1) English dataset: mean, 73.695273; median, 73.0; min/max, 1/199. (2) Hungarian dataset: mean, 76.833792; median, 75.0; min/max, 1/199. (3) Japanese dataset: mean, 34.873541; median, 38.0; min/max, 1/162. (4) Full (EN-HU-JP) dataset: mean, 65.302173; median, 57.0; min/max, 1/199.

Figure 9 shows a visualization of the distribution of text length in the English, Hungarian, and Japanese datasets for experiment 3 (Experiment1_50M_enhujp_v2_2), where the horizontal axis represents the text length and the vertical axis represents the frequency on each subchart. The mean, median, minimum, and maximum values for the filtered text lengths and for the overall text length without considering the language are as follows. (1) English dataset: mean, 76.840730; median, 77.0; min/max, 3/199. (2) Hungarian dataset: mean, 79.699119; median, 78.0; min/max, 3/199. (3) Japanese dataset: mean, 35.632283; median, 39.0; min/max, 3/162. (4) Full (EN-HU-JP) dataset: mean, 67.751759; median, 59.0; min/max, 3/199.

The approach was validated on different datasets. The result and experiments are presented in [47].

6. Discussion

Our primary objective was to ensure that the dataset adequately represents the unique characteristics of each language (Japanese, English, and Hungarian), including the alphabets, most frequently used characters, writing styles, and common vocabulary. This allows the OCR system to learn and recognize the distinctions and combinations between or among language scripts effectively. The dataset that was used in the experiments contains high-quality, clean, and accurate text samples from well-known and recent sources to train the OCR system effectively. This ensured that the model learned to recognize characters and words—as well as hybrid languages—accurately, improving its overall performance in real-time OCR tasks. The training dataset prepared incorporated samples with various degrees of noise, such as occlusions, distortions, or low-quality images, to train the OCR system to be robust and tolerant to noise in real-world scenarios. During experiments and the dataset preparation workflow optimization process, we included a wide variety of text samples, such as different fonts, writing styles, and contexts, which helped the OCR system generalize better and adapt to various scenarios. This diversity in data enables the model to handle a wide range of text inputs from different sources. We established that a "balanced dataset" is essential to prevent the model from being biased towards a particular language or character set. As a future outcome, this involves the "equal" representation of each language and the mitigation of class imbalances, enabling the OCR system to recognize characters and words from all three languages equally well. It was ensured that the training dataset included samples that reflect the contextual variations of each language and their variations (e.g., English acronyms or technical terms in Hungarian or English phrases). This allows the OCR system to understand and interpret texts or phrases within different contexts, such as formal, informal, technical, or colloquial settings. We incorporated domain-specific data for each language to support the OCR system in specialized fields or use cases, such as legal, IT, medical, engineering, or technical contexts. This adaptation allows the model to recognize and interpret domain-specific vocabulary and jargon more accurately. In order to allow for further expandability, a specific dataset was designed to allow for easy integration of additional languages or new data samples in the future or to be ported to a HPC environment. This facilitates the ongoing improvement and expansion of the OCR system to cater to new requirements and evolving language trends.

Limitations of the Current Approach

While this research endeavors to address significant challenges in multilanguage OCR and synthesized languages for collaborative on-screen OCR solutions (especially in health care and remote diagnosis), there were localized limitations that underscore the complexity of the task. The specific constraints that necessitate thorough examination are as follows:

1. *Synthesized Languages*: Investigating synthesized languages (like health language, legal language, and IT language), which may not have well-established linguistic resources or standardized writing systems, introduces an additional layer of complexity. The availability of reliable linguistic references and textual corpora for such languages may be limited, potentially hampering the completeness and authenticity of the synthetic training data.
2. *Computational Resources*: This research involves computationally intensive tasks, including the generation of synthetic data and the training of deep learning models. Depending on the scale of the dataset and the complexity of the models, substantial computational resources or a high-performance computing (HPC) environment is required. Access to such resources may impose constraints, particularly for researchers with limited access to HPC environments.
3. *Generalization*: while this research seeks to optimize training datasets and models for multilanguage OCR, the extent to which the findings can be generalized to various

OCR applications and domains may be limited. Factors such as document types, writing styles, and use cases may influence the transferability of the results.

4. *Evaluation Challenges*: Establishing robust, metadata-based evaluation methodologies and benchmarks for multilanguage OCR systems that consider the complexities of synthesized languages is a formidable task. The creation of evaluation frameworks that adequately reflect real-world scenarios (like cross-domain or field cooperation) and effectively assess system performance across multiple languages is an ongoing challenge.
5. *Data Annotation*: The generation of high-quality, linguistically accurate annotations for synthetic training datasets is a resource-intensive and time-consuming endeavor. The practicality of creating comprehensive annotations that encompass diverse linguistic features across multiple languages and/or domains (law, IT, health care, sports, and industry), especially for context-sensitive and synthesized languages, may present significant logistical challenges, even in HPC environments.
6. *Language Diversity Coverage*: One prominent limitation pertains to the diverse nature of the languages involved. While this research aims to address multilingual challenges, each language possesses unique characteristics, scripts, and linguistic nuances. Consequently, the development of a single, universally applicable methodology for training dataset preparation may be challenging. The need for language-specific considerations may introduce complexity and complicate the generalization of findings across languages. Icluding Japanese language and synthesized language alongside Latin and Anglo-Saxon languages can limit the vocabulary to 5000–6500 characters if the language is limited to the IT domain, but in the case of the health domain, the vocabulary will have include than 10,000 characters.

Acknowledging these limitations is essential to provide a balanced perspective and guide future research efforts in the demanding and dynamic field of OCR technology.

7. Future Development

In future phases, we will expand the experiments to an HPC environment (Komondor [57]), where the proposed approach will be used for federated learning and distributed data processing combined with domain adaptation and unsupervised learning methods using dynamic data sampling and the continuous model improvement approach.

8. Open Challenges and Problems

Generating synthetic training datasets for the training of a mixed natural (e.g., Japanese, English, and Hungarian) or synthesized multilanguage (e.g., industry, health, law, and IT language) OCR system presents various challenges and problems, both using "classic AI infrastructure" and in an HPC (high-performance computing) cluster environment. Table 6 shows a comparison of the challenges and considerations between the two scenarios and the reasons why we propose the application of multilingual OCR approaches to HPC environments.

In conclusion, while both classic AI infrastructure and HPC clusters can be used for synthetic training dataset generation for multilanguage OCR, HPC environments offer significant advantages in terms of computational power, scalability, and resource management. However, challenges related to data quality, labeling, variability, and resource availability remain common in both scenarios and require careful consideration in the dataset generation process.

Table 6. Comparison of open challenges comparison: classic AI infrastructure vs. an HPC environment.

Topic	Classic AI Infrastructure	HPC Cluster Environment
Computational Power/Resources	In classic AI infrastructure, the user is limited by the computational resources available on a single machine or in a small cluster. Generating synthetic training data for multiple languages with a large dataset can be computationally intensive and time-consuming.	HPC clusters offer significantly more computational power, allowing the user to efficiently generate larger and more complex synthetic datasets for multilanguage OCR.
Scalability	Classic AI infrastructure may lack the scalability required to efficiently simultaneously generate synthetic datasets multiple languages. The user may need to create separate work flows for each language, which can be inefficient.	HPC clusters are designed for parallel processing, which means the user can distribute the workload across multiple nodes, enabling simultaneous dataset generation for multiple languages. This enhances efficiency and reduces time requirements.
Data Handling/Volume	Managing and storing large volumes of synthetic data for multiple languages can be challenging in a traditional setup. The user might encounter storage limitations and data management issues.	HPC environments provide advanced data storage and management capabilities. The handling and storage of large volumes of synthetic data are streamlined.
Parallelism/Distributed Computing	Implementing parallelism for data generation is more complex in classic infrastructure, making it harder to exploit multicore CPUs or GPUs efficiently.	HPC clusters are well-suited for distributed computing, enabling the user to distribute the dataset generation tasks across multiple nodes, which is especially beneficial for the generation of synthetic datasets for multiple languages.
Resource Management	—	HPC environments offer sophisticated resource management tools and job schedulers, allowing the user to optimize resource allocation and utilization for dataset generation tasks.

9. Conclusions

The optimal dataset can handle code switching; and is robust to noise; is capable of domain adaptation, transfer learning, and handling multimodal data, scalability, as well as benchmarking and evaluation. These factors contribute to the development of accurate, efficient, and versatile OCR systems capable of handling multiple languages and various real-world scenarios.

The most important findings of our experiments can be summarized as follows:

1. Preparing an optimal dataset directly contributes to improved model performance by providing diverse, high-quality, and representative samples for each language, enabling the OCR systems to learn and generalize better; however, this process needs continuous process control.
2. Preparing a specific dataset that includes noisy samples, such as low-quality images, occlusions, or distortions, ensures that the OCR system is robust and can perform accurately even in challenging real-world conditions.
3. By incorporating multiple languages or hybrid languages, the training dataset enables the OCR system to cater to a broader audience, increasing its applicability and utility in various real-world scenarios. This model could be interesting not only for multilanguage use cases but also for training with technical and/or domain-specific languages. Moreover, by incorporating domain-specific data and contexts, the OCR system can be tailored to different industries or use cases, improving its performance in specialized fields such as the legal, medical, or technical domains.
4. Preparing a dataset that includes multimodal information such as images or audio helps the OCR system leverage additional contextual information to enhance its understanding and interpretation of the text. This process may make the preparation stage more complicated, but the results will compensate for that.
5. The dataset should include variations in writing styles, dialects, and colloquialisms for each language, ensuring that the OCR system is adaptable and can handle various linguistic nuances.
6. Multilingual societies frequently engage in “code switching”, the act of switching between two or more languages within a single conversation or document. An optimal dataset should account for code-switching scenarios, allowing the OCR system to seamlessly recognize and process text that contains multiple languages.
7. Leveraging pretrained multilingual models or incorporating cross-lingual pretraining enables the OCR system to learn from the shared knowledge of multiple languages, resulting in better generalization and potentially improved performance for low-resource languages.

The key finding is that ML models developed following the methods outlined in this research perform consistently well and accurately. Multilingual/hybrid texts were accurately detected by the ML models. The highest precision scores were attained in the following experiments: "30M_enhujp_v2_4", with 98.76% "30M_enhujp_v2_1", with 95.87%; and "50M_enhujp_v2_2" with 94% accuracy.

In summary, the proposed methodology justifies the notion that optimal dataset preparation in real-time multilanguage or hybrid language training requires a well-defined pipeline with domain expertise. Following the proposed methodology will result in enhanced model performance, providing broader language coverage and adaptability to language variations at the same time. Using the methodology presented in this paper will provide more flexibility for complex OCR projects, as well as a visual dashboard-controlled process for a better outcome and for quick dataset adjustments.

Author Contributions: Conceptualization, A.B. and S.M.S.; methodology, A.B. and L.S.; software, A.B.; validation, A.B. and L.S.; formal analysis, A.B. and S.M.S.; investigation, A.B.; resources, A.B. and L.S.; data curation, A.B.; writing—original draft preparation, A.B.; writing—review and editing, A.B. and L.S.; visualization, A.B. and L.S.; supervision, S.M.S. and L.S.; project administration, A.B.; funding acquisition, A.B. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by ITware, Hungary. The work of A.B was supported by the University of Malaga. The work of S.M.S. was supported by the Department of Electrical Engineering and Information Technology of George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures. The work of L.S. was supported by the Consolidator Excellence Researcher Program of Óbuda University, Budapest, Hungary.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available upon request. The data are not publicly available due to privacy. Figures and the results of this study can be found at https://doi.org/10.6084/m9.figshare.22151651.v2 (accessed on 26 October 2023).

Acknowledgments: The HPC feasibility and comparison test was supported by the new Hungarian supercomputer Komondor, operated by the Governmental Agency for IT Development. The authors offer special thanks to Antonio I. Cuesta-Vargas, Jaime Martin-Martin, and the University of Malaga (UMA) for continuous support, technical guidance, and mentoring.

Conflicts of Interest: The authors declare that this study received funding from ITware. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
OCR	Optical character recognition
GANs	Generative adversarial networks
VAEs	Variational autoencoders
CNN	convolutional neural network
RNN	Recurrent neural network
CRNN	Convolutional recurrent neural network
SVTR	Single Visual model for scene Text Recognition
IT	Information technology
MHSA	Multihead self-attention
MHA	Multihead attention
UX	User experience
HLP	High-performance computing
PC	Personal computer
CPU	Central processing unit

GB	Gigabyte
GPU	Graphics processing unit
VGA	Video graphics array

References

1. Biró, A.; Jánosi-Rancz, K.T.; Szilágyi, L.; Cuesta-Vargas, A.I.; Martín-Martín, J.; Szilágyi, S.M. Visual Object Detection with DETR to Support Video-Diagnosis Using Conference Tools. *Appl. Sci.* **2022**, *12*, 5977. [CrossRef]
2. Benis, A.; Grosjean, J.; Billey, K.; Martins, J.; Dornauer, V.; Crișan-Vida, M.; Hackl, W.; Stoicu-Tivadar, L.; Darmoni, S. Medical Informatics and Digital Health Multilingual Ontology (MIMO): A tool to improve international collaborations. *Int. J. Med. Inform.* **2022**, *167*, 104860. [CrossRef] [PubMed]
3. Shah, S.R.; Kaushik, A.; Sharma, S.; Shah, J. Opinion-Mining on Marglish and Devanagari Comments of YouTube Cookery Channels Using Parametric and Non-Parametric Learning Models. *Big Data Cogn. Comput.* **2020**, *4*, 3. [CrossRef]
4. Shah, S.R.; Kaushik, A. Sentiment Analysis on Indian Indigenous Languages: A Review on Multilingual Opinion Mining. *arXiv* **2019**, arXiv:1911.12848. [CrossRef]
5. Pathak, K.; Saraf, S.; Wagh, S.; Vishwanath, D. OCR Studymate. *Int. J. Res. Appl. Sci. Eng. Technol.* **2022**, *10*, 2241–2246. [CrossRef]
6. Nuchkrua, T.; Leephakpreeda, T. Novel Compliant Control of a Pneumatic Artificial Muscle Driven by Hydrogen Pressure Under a Varying Environment. *IEEE Trans. Ind. Electron.* **2022**, *69*, 7120–7129. [CrossRef]
7. Sharma, P. Advancements in OCR: A Deep Learning Algorithm for Enhanced Text Recognition. *Int. J. Invent. Eng. Sci.* **2023**, *10*, 1–7. [CrossRef]
8. Subedi, B.; Yunusov, J.; Gaybulayev, A.; Kim, T.H. Development of a low-cost industrial OCR system with an end-to-end deeplearning technology. *IEMEK J. Embed. Syst. Appl.* **2020**, *15*, 51–60.
9. Chen, Y.H.; Zhou, Y.L. Enhancing OCR Performance through Post-OCR Models: Adopting Glyph Embedding for Improved Correction. *arXiv* **2023**, arXiv:2308.15262.
10. Nieminen, H.; Kuosmanen, L.; Bond, R.; Vartiainen, A.-K.; Mulvenna, M.; Potts, C.; Kostenius, C. Coproducing multilingual conversational scripts for a mental wellbeing chatbot-where healthcare domain experts become chatbot designers. *Eur. Psychiatry* **2022**, *65*, S293. [CrossRef]
11. Mao, A.X.Q.; Thakkar, I. Lost in Translation: The Vital Role of Medical Translation in Global Medical Communication. *AMWA J.* **2023**, *38*, 3.
12. Yilmaz, B.; Korn, R. Understanding the mathematical background of Generative Adversarial Networks (GANs). *Math. Model. Numer. Simul. Appl.* **2023**, *3*, 234–255. [CrossRef]
13. Moghaddam, M.M.; Boroomand, B.; Jalali, M.; Zareian, A.; Daeijavad, A.; Manshaei, M.H.; Krunz, M. Games of GANs: Game-theoretical models for generative adversarial networks. *Artif. Intell. Rev.* **2023**, *56*, 9771–9807. [CrossRef]
14. Singh, A.; Ogunfunmi, T. An Overview of Variational Autoencoders for Source Separation, Finance, and Bio-Signal Applications. *Entropy* **2021**, *24*, 55. [CrossRef] [PubMed]
15. Ebrahimnejad, J.; Naghsh, A.; Pourghasem, H. A robust watermarking approach against high-density salt and pepper noise (RWSPN) to enhance medical image security. *IET Image Proc.* **2023**. [CrossRef]
16. Gao, J.Q.; Li, L.; Ren, X.; Chen, Q.; Abdul-Abbass, Y. An effective method for salt and pepper noise removal based on algebra and fuzzy logic function. *Multim. Tools Appl.* **2023**. [CrossRef]
17. Muthmainnah. Optimized the Performance of Super Resolution Images by Salt and pepper Noise Removal based on a Modified Trimmed Median Filter. *Wasit J. Comput. Math. Sci.* **2023**, *2*, 107–115. [CrossRef]
18. Kumain, S.; Kumar, K. Quantifying Salt and Pepper Noise Using Deep Convolutional Neural Network. *J. Inst. Eng. Ser. B* **2022**, *103*, 1293–1303. [CrossRef]
19. Tian, Y.; Wu, S.; Zeng, J.; Gao M. PaddleOCR—An Elegant And Modular Architecture. DESOSA2021, 15 March 2021. Available online: https://2021.desosa.nl/projects/paddleocr/posts/paddleocr-e2/ (accessed on 10 October 2023).
20. Monteiro, G.; Camelo, L.; Aquino, G.; Fernandes, R.; Gomes, R.; Printes, A.; Gondres, I.; Silva, H.; Parente de Oliveira, J.; Figueiredo, C. A Comprehensive Framework for Industrial Sticker Information Recognition Using Advanced OCR and Object Detection Techniques. *Appl. Sci.* **2023**, *13*, 7320. [CrossRef]
21. Du, Y.N.; Li, C.X.; Guo, R.Y.; Cui, C.; Liu, W.W.; Zhou, J.; Lu, B.; Yang, Y.H.; Liu, Q.W.; Hu, X.G.; et al. PP-OCRv2:Bag of tricks for ultra lightweightOCR system. *arXiv* **2021**, arXiv:2109.03144.
22. Guo, Q.; Zhang, C.; Zhang, S.; Lu, J. Multi-model query languages: Taming the variety of big data. *Distr. Paral. Databases.* **2023**. [CrossRef]
23. Jain, P.; Taneja, K.; Taneja, H. Which OCR toolset is good and why? A comparative study. *Kuwait J. Sci.* **2021**, *48*, 1–12. [CrossRef]
24. Jain, P.; Kumar, V.; Samuel, J.; Singh, S.; Mannepalli, A.; Anderson, R. Artificially Intelligent Readers: An Adaptive Framework for Original Handwritten Numerical Digits Recognition with OCR Methods. *Information* **2023**, *14*, 305. [CrossRef]
25. Yoshimura, M.; Otsuka, J.; Irie, A.; Ohashi, T. Rawgment: Noise-Accounted RAW Augmentation Enables Recognition in a Wide Variety of Environments. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14007–14017.
26. Jadhav, G.; Jada, S.Y.; Triveni, R.; Khan, S.I. FPGA based Edge Detection using Sobel Filter. *Int. J. Res. Appl. Sci. Eng. Technol.* **2021**, *9*, 145–147.

27. Le, H.; Kim, S.H.; Na, I.; Do, Y.; Park, S.C.; Jeong, S.H. Automatic Generation of Training Character Samples for OCR Systems. *Int. J. Contents* **2012**, *8*, 83–93. [CrossRef]
28. Xu, Y.J. An Adaptive Learning System for English Vocabulary Using Machine Learning. *Mobile Inform. Syst.* **2022**, *2022*, 3501494. [CrossRef]
29. März, L.; Schweter, S.; Poerner, N.; Roth, B.; Schütze, H. Data Centric Domain Adaptation for Historical Text with OCR Errors. In *International Conference on Document Analysis and Recognition (ICDAR)*; Springer: Cham, Switzerland; Berlin/Heidelberg, Germany, 2021; Volume 12822, pp. 748–761.
30. Blevins, T.; Gonen, H.; Zettlemoyer, L. Analyzing the Mono- and Cross-Lingual Pretraining Dynamics of Multilingual Language Models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 3575–3590.
31. Nowakowski, K.; Ptaszynski, M.; Murasaki, K.; Nieuwazny, J. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Inform. Proc. Manag.* **2023**, *60*, 103148. [CrossRef]
32. Faizullah, S.; Ayub, M.S.; Hussain, S.; Khan, M.A. A Survey of OCR in Arabic Language: Applications, Techniques, and Challenges. *Appl. Sci.* **2023**, *13*, 4584. [CrossRef]
33. Spruck, A.; Hawesch, M.; Maier, A.; Riess, C.; Seiler, J.; Kau, A. 3D Rendering Framework for Data Augmentation in Optical Character Recognition. In Proceedings of the International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 15–16 July 2021.
34. Milyaev, S.; Barinova, O.; Novikova, T.; Kohli, P.; Lempitsky, V. Fast and accurate scene text understanding with image binarization and off-the-shelf OCR. *Int. J. Doc. Anal. Recogn. (IJDAR)* **2015**, *18*, 169–182. [CrossRef]
35. Englmeier, T.; Fink, F.; Springmann, U.; Schulz, K. Optimizing the Training of Models for Automated Post-Correction of Arbitrary OCR-ed Historical Texts. *J. Lang. Technol. Comput. Linguist.* **2022**, *35*, 1–27. [CrossRef]
36. McKinzie, R.; Cheng, J.; Shankar, V.; Yang, Y.F.; Shlens, J.; Toshev, A. On robustness in multimodal learning. *arXiv* **2023**, arXiv:2304.04385.
37. Sansowa, R.; Abraham, V.; Patel, M.; Gajjar, R. OCR for Devanagari Script Using a Deep Hybrid CNN-RNN Network. *Lect. Notes Electr. Eng.* **2022**, *952*, 263–274.
38. Okamoto, S.; Jin'no, K. A study on the role of latent variables in the encoder-decoder model using image datasets. *Nonlin. Theor. Its Appl. (IEICE)* **2023**, *14*, 652–676. [CrossRef]
39. Zhang, M.; Duan, Y.; Song, W.; Mei, H.; He, Q. An Effective Hyperspectral Image Classification Network Based on Multi-Head Self-Attention and Spectral-Coordinate Attention. *J. Imag.* **2023**, *9*, 141. [CrossRef]
40. Sang, D.V.; Cuong, L.T.B. Improving CRNN with EfficientNet-like feature extractor and multi-head attention for text recognition. In Proceedings of the 10th International Symposium on Information and Communication Technology (SoICT), Hanoi, Vietnam, 4–6 December 2019; pp. 285–290.
41. Jaiswal, K.; Suneja, A.; Kumar, A.; Ladha, A.; Mishra, N. Preprocessing Low Quality Handwritten Documents for OCR Models. *Int. J. Res. Appl. Sci. Eng. Technol.* **2023**, *11*, 2980–2985. [CrossRef]
42. Du, Y.; Chen, Z.; Jia, C.; Yin, X.; Zheng, T.; Li, C.; Du, Y.; Jiang, Y.-G. SVTR: Scene Text Recognition with a Single Visual Model. In Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI), Vienna, Austria, 23–29 July 2022; pp. 867–873.
43. Goel, P.; Bansal, S. Comprehensive and Systematic Review of Various Feature Extraction Techniques for Vernacular Languages. In Proceedings of the Innovations in Bio-Inspired Computing and Applications (IBICA 2022), Online, 15–17 December 2022; pp. 350–362.
44. Zulkifli, M.K.N.; Daud, P.; Mohamad, N. Multi Language Recognition Translator App Design Using Optical Character Recognition (OCR) and Convolutional Neural Network (CNN). In Proceedings of the International Conference on Data Science and Emerging Technologies (DaSET 2022), Online, 20–21 December 2022; pp. 103–116.
45. Biten, A.F.; Tito, R.; Gomez, L.; Valveny, E.; Karatzas D. OCR-IDL: OCR Annotations for Industry Document Library Dataset. In *European Conference on Computer Vision (ECCV 2022)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 13804, pp. 241–252.
46. Feng, C.; Gong, M.; Deussen, O. A Balanced-Partitioning Treemapping Method for Digital Hierarchical Dataset. *Virt. Real. Intell. Hardw.* **2022**, *4*, 342–358. [CrossRef]
47. Biró, A.; Cuesta-Vargas, A.I.; Martín-Martín, J.; Szilágyi, L.; Szilágyi, S.M. Synthetized Multilanguage OCR Using CRNN and SVTR Models for Realtime Collaborative Tools. *Appl. Sci.* **2023**, *13*, 4419. [CrossRef]
48. Kim, G.; Hong, T.; Yim, M.; Nam, J.; Park, J.; Yim, J.; Hwang, W.; Yun, S.; Han, D.; Park, S. OCR-Free Document Understanding Transformer. In Proceedings of the European Conference on Computer Vision (ECCV 2022), Tel Aviv, Israel, 23–27 October 2022.
49. Akasapu, A.; Sailaja, V.; Prasad, G. Implementation of Sobel filter using CUDA. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1045*, 012016. [CrossRef]
50. Eixelberger, T.; Wolkenstein, G.; Hackner, R.; Bruns, V.; Mühldorfer, S.; Geissler, U.; Belle, S.; Wittenberg, T. YOLO networks for polyp detection: A human-in-the-loop training approach. *Curr. Dir. Biomed. Eng.* **2022**, *8*, 277–280. [CrossRef]
51. He, Y.; Yuan, J.L.; Li, L. Enhancing RNN Based OCR by Transductive Transfer Learning From Text to Images. *AAAI Conf. Artif. Intell.* **2018**, *32*, 8083–8084. [CrossRef]

52. Dhanya, D.; Ramakrishnan, A.G. Optimal feature extraction for bilingual OCR. *Lect. Notes Comp. Sci.* **2002**, *2423*, 25–36.
53. Kim, J.; Huh, J.; Park, I.; Bak, J.; Kim, D.; Lee, S. Small Object Detection in Infrared Images: Learning from Imbalanced Cross-Domain Data via Domain Adaptation. *Appl. Sci.* **2022**, *12*, 11201. [CrossRef]
54. Rane, T.; Bhatt, A. A Deep Learning-Based Regression Scheme for Angle Estimation in Image Dataset. In Proceedings of the International Conference on Recent Trends in Image Processing and Pattern Recognition (RTIP2R 2022), Kingsville, TX, USA, 1–2 December 2023; pp. 282–296.
55. Xin, F.F.; Zhang, H.P.; Pan, H.Q. Hybrid dilated multilayer faster RCNN for object detection. *Vis. Comput.* **2023**. [CrossRef]
56. PP-OCR. Available online: https://github.com/PaddlePaddle/PaddleOCR/blob/v2.5.0/doc/doc_en/ppocr_introduction_en.md#pp-ocrv3 (accessed on 6 December 2023).
57. Komondor, One of the Greenest Supercomputers in the World, HPC Competence Center. Available online: https://hpc.kifu.hu/en/komondor-one-of-the-greenest-supercomputers-in-the-world (accessed on 13 October 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.