On the 17th of October 2023, we mark a significant moment in our project's journey. We are pleased to announce the start of our data collection phase, a crucial step in developing our "Handwritten Questions Grading System." During this phase, we have discovered and obtained databases that contain handwritten text, which promises to greatly enhance our project.

These databases are rich sources of handwritten content, offering valuable insights and resources that will help us improve our grading system. With these collections at our disposal, we are well-prepared to use technology like machine learning and natural language processing to make the grading process faster, more accurate, and less biased.

As we embark on this data collection journey, we anticipate the potential these databases hold. They are not just collections of text but tools that can help us transform the way we evaluate educational assessments. This exciting endeavor brings us closer to reshaping the education landscape to be more efficient, fair, and objective.

Database Description: IAM Handwriting Database

The initial database at our disposal is the esteemed IAM Handwriting Database, a repository of handwritten English sentences derived from the venerable Lancaster-Oslo/Bergen (LOB) corpus. The LOB corpus, renowned for its expansive textual collection encompassing approximately one-million-word instances, serves as the foundational source for this database. Within its confines, one finds 1,066 forms, thoughtfully created by nearly 400 distinct writers, and an impressive tally of 82,227-word instances culled from a vocabulary of 10,841 words.

This database is predominantly constituted of entire English sentences, making it a repository of immense value for a wide array of tasks related to handwriting recognition. Its true potential, however, emerges when harnessed for recognition tasks that necessitate linguistic knowledge beyond the lexicon level, as this invaluable insight can be systematically extracted from the corpus upon which it is built. Furthermore, this database has been endowed with a suite of image-processing procedures, thoughtfully designed to extract handwritten text from the forms and to meticulously segment the text into lines and words.

The IAM Handwriting Database assumes a pivotal role in the training and validation of handwritten text recognition systems, and it is also a valuable asset for the execution of writer identification and verification experiments. It is noteworthy that this database, in adherence to principles of open access and dissemination, is publicly accessible and available without charge for non-commercial research endeavors. In the context of our project, this acquisition marks a significant enrichment, affording us a robust foundation for the development of our automated grading system.

IAM Handwriting Database 3.0

Overview

The IAM Handwriting Database, version 3.0, stands as a comprehensive repository of handwritten English text, purposefully designed to facilitate the training and testing of handwritten text recognition systems, as well as to support endeavors in writer identification and verification. It is noteworthy that the inception of this database can be traced back to its initial publication at the International Conference on Document Analysis and Recognition (ICDAR) in 1999, marking the genesis of an indispensable resource for the research community.

Over the years, the IAM Handwriting Database has played a pivotal role in advancing the field of handwritten text recognition. This contribution culminated in the development of an HMM-based recognition system for handwritten sentences, the results of which were disseminated at the International Conference on Pattern Recognition (ICPR) in 2000. Moreover, the segmentation scheme employed in the second iteration of the database was meticulously documented and shared with the scientific community at ICPR in 2002. The ongoing significance of the IAM database is underscored by its continued utilization in our own research endeavors, which is substantiated by a portfolio of publications.

Data Characteristics

The IAM Handwriting Database 3.0 boasts the following notable characteristics:

- **Contributions:** A total of 657 writers have generously contributed samples of their handwriting, augmenting the diversity and richness of the database.

- **Scanned Text Pages:** The database encompasses 1,539 pages of meticulously scanned text, offering an extensive corpus for research and experimentation.

- **Isolated and Labeled Sentences:** A substantial collection of 5,685 isolated sentences, thoughtfully labeled for clarity and accuracy.

- **Isolated and Labeled Text Lines:** An array of 13,353 isolated and labeled text lines, instrumental for tasks requiring line-level recognition.

- **Isolated and Labeled Words:** The database comprises an impressive array of 115,320 isolated and labeled words, carefully extracted from scanned pages using an automated segmentation scheme, which has undergone meticulous manual verification. This segmentation scheme, a product of our institution's research, has been rigorously developed.

The entire database is provided in the form of PNG image files, with accompanying XML meta-information thoughtfully embedded within these images. The XML meta-information includes segmentation details and an array of estimated parameters, derived from the preprocessing procedures documented in [1]. This meta-information is available in XML format and adheres to the XML file and XML file format (DTD) specifications.

Database Access

For those seeking access to the IAM Handwriting Database, this invaluable resource is readily available for download through the following link: Download the IAM Handwriting Database. This link facilitates direct access to the database, adhering to open-access principles and providing an opportunity for non-commercial research initiatives to explore its rich content. This database serves as a pivotal asset for research in handwriting recognition, text analysis, and linguistics, and its accessibility aligns with the highest standards of research dissemination.

[1] M. Zimmermann and H. Bunke. Automatic Segmentation of the IAM Off-line Database for Handwritten English Text. In Proc. of the 16th Int. Conf. on Pattern Recognition, Volume 4, pages 35 - 39, 2000.

Database Description: CROHME - Competition on Recognition of Online Handwritten Mathematical Expressions

The CROHME dataset stands as a monumental collection, encompassing over 12,000 handwritten mathematical expressions contributed by hundreds of writers hailing from various corners of the globe. This amalgamation consolidates datasets from four previous CROHME competitions while infusing new resources into the mix. Writers were tasked with transcribing printed expressions sourced from a carefully curated corpus. This corpus, meticulously assembled to encapsulate the diversity of tasks, draws from existing mathematical corpora and expressions found within Wikipedia pages. An array of devices was employed, ranging from diverse digital pen technologies to white-board input devices and tablets featuring sensitive screens, each operating at varying scales and resolutions. Notably, the dataset exclusively provides an on-line signal, offering a distinctive focus on dynamic data.

Dataset Characteristics and Origin:

- **Expressions:** The dataset comprises more than 12,000 expressions, bearing the unique marks of hundreds of contributors spanning diverse cultural backgrounds. These expressions were diligently copied from a corpus of mathematical expressions.

- **Device Variability:** The dataset exhibits a spectrum of writing instruments and technologies employed, encompassing different digital pen technologies, white-board input devices, and tablets with sensitive screens. This diversity in devices translates to a range of scales and resolutions in the captured data.

Competition Evolution:

In the latest iteration, CROHME 2013, the test section introduces entirely original content, while the training section leverages five existing datasets:

1. MathBrush (University of Waterloo)
2. HAMEX (University of Nantes)
3. MfrDB (Czech Technical University)
4. ExpressMatch (University of Sao Paulo)
5. the KAIST dataset

Additionally, participants from the 2012 competition have graciously provided their recognized expressions for the 2012 test section. This affords researchers the opportunity to delve into aspects such as decision fusion or evaluation metrics.

Metadata and Ground Truth Data:

The CROHME dataset meticulously incorporates segmentation, labeling, and layout details for each mathematical expression, adhering to the INKML and MATHML standards. Within an InkML file, three crucial components are encapsulated:

- The ink, manifested as a series of traces constituted by points.
- Symbol-level ground truth, comprising segmentation and labeling information for each symbol within the expression.
- Expression-level ground truth, delineating the MathML structure of the expression.

Both levels of ground truth data, symbol-level and expression-level, are meticulously entered by hand. Furthermore, the file incorporates general information including channels (X and Y), writer specifics (identification, handedness, age, gender, etc., if available), LaTeX ground truth (independent of ink, thus easily renderable), and the unique identification code of the ink (UI), among other details. The InkML format serves as the conduit linking digital ink, symbol segmentation, and MathML representation, facilitating a comprehensive understanding of the data.

Recognized Expressions:

The recognized expressions denote the outcomes of the competing recognition systems. They adhere to the same InkML format, sans the ink information, retaining only segmentation, labeling, and MathML structure.

The CROHME dataset emerges as a foundational resource for the recognition of online handwritten mathematical expressions, offering unparalleled depth in data and ground truth information. Its extensive scope and meticulous curation make it an invaluable asset for researchers and practitioners alike.

Accessing and Downloading the CROHME 2023 Dataset

The CROHME 2023 dataset constitutes a comprehensive repository of data, encompassing contributions from previous CROHME competitions, the OffRaSHME competition, new image samples, and forthcoming bimodal data. In addition to the online equation data, this dataset also furnishes corresponding images, generated from either online ink or scanned from the original physical pages. Notably, the equations from OffRaSHME are exclusively in an off-line format.

For the convenience of researchers and practitioners, the complete dataset package is available for download via the following link: CROHME 2023 Dataset.

To initiate the download process:

1. Access the provided link by clicking on it.
2. Ensure that your device has adequate storage capacity to accommodate the dataset, as it may be substantial due to the extensive nature of the data.

Kindly note that the 2019 test set is primarily intended for reference purposes, serving as a benchmark to compare system performance with previous competition results. It is not recommended for utilization in the training or validation of current systems.

The CROHME 2023 dataset serves as a valuable resource for researchers, offering an extensive and meticulously curated collection of data and ground truth information for the recognition of online handwritten mathematical expressions. Its inclusion of diverse data sources and formats makes it an indispensable asset for those engaged in research and practice in this field.

Database Description: HME100K - Handwritten Mathematical Expression Dataset

As we delve deeper into the realm of our data collection phase, we are honored to introduce the HME100K dataset, a monumental addition to our resources. This large-scale and real scene dataset is poised to serve as a cornerstone for the evaluation of handwritten mathematical expression recognition tasks, marking a significant stride in our project.

Image Collection:

The HME100K dataset bears testimony to its real-world relevance, meticulously curated from the collective efforts of tens of thousands of writers. Their dedicated endeavors have left an indelible mark on this dataset, as they transcribed mathematical expressions onto paper, which were then thoughtfully uploaded to an internet application, forming an extensive and authentic collection.

Dataset Composition:

Our team is privileged to have access to the HME100K dataset, a compendium of 99,109 images thoughtfully segregated into 74,502 training images and 24,607 testing images. This expansive dataset stands as a testament to its diversity, encompassing a remarkable 245 distinct symbol classes. Notably, the HME100K dataset presents a tenfold increase in data size when compared to its predecessor, the CRHOME dataset. This monumental expansion underlines the magnitude of contributions from tens of thousands of writers via an internet application, adding layers of complexity and richness to the dataset.

Accessibility:

To gain access to this invaluable resource, we invite you to visit the official website: HME100K Dataset. This repository, with a file size of approximately 695.77MB, provides a wealth of data for research and experimentation, reinforcing its status as an indispensable resource for our project and the broader community of researchers.

The HME100K dataset represents a significant step forward in the realm of handwritten mathematical expression recognition. Its real-world applicability, scale, and the diversity of expressions transcribed by tens of thousands of contributors signify its pivotal role in our project. We are committed to leveraging this dataset to advance the boundaries of handwritten mathematical expression recognition, and we anticipate its transformative impact on our project's goals and objectives.

Database Description: Bentham Dataset R0

In our quest for comprehensive datasets, we are privileged to introduce the Bentham Dataset R0, a reservoir of invaluable works on law and moral philosophy penned by the distinguished philosopher Jeremy Bentham. This collection stands as a testament to Bentham's enduring influence on the realms of jurisprudence and ethical philosophy.

Dataset Composition:

The Bentham collection comprises a meticulous compilation of images capturing the essence of Jeremy Bentham's writings. These documents, bearing the imprint of his philosophical musings, are made available for research purposes, reflecting the ethos of open scholarship and collaborative inquiry.

Components:

This dataset is thoughtfully divided into two integral components: the images and the Ground Truth (GT). The Ground Truth encompasses vital information regarding the layout and line-level transcription of each image, meticulously organized in the PAGE format. It is imperative to note that both components must be procured separately, and a detailed exposition of the dataset's organizational structure is provided within each part.

Bentham Manuscripts:

The Bentham manuscripts represent an extensive body of documents authored by Jeremy Bentham himself, a luminary in the realm of English philosophy and social reform. A remarkable facet of this dataset lies in its genesis from the Transcribe Bentham initiative, a testament to the power of collective effort. Through this public web platform, volunteers have transcribed over 6,000 documents, encompassing an impressive 25,000 pages, thus breathing life into this invaluable collection.

Accessibility:

For those keen on accessing this trove of knowledge, we extend an invitation to visit the official repository: Bentham Dataset R0. This resource, marked by its accessibility and dedication to advancing research, represents an invaluable asset for our project and the broader academic community.

The Bentham Dataset R0 is emblematic of our commitment to drawing upon diverse and meticulously curated resources in our pursuit of a revolutionary Handwritten Questions Grading System. This dataset, steeped in the intellectual legacy of Jeremy Bentham, promises to play a pivotal role in shaping the trajectory of our project. We anticipate that its profound insights and scholarly contributions will significantly bolster our endeavors in the field of educational assessment.

Database Description: Rimes Dataset

In our diligent pursuit of datasets to enhance our project, we recognize the significance of the Rimes dataset. Although it may not bear the same weight as other datasets, given its historical nature, it still holds a valuable place in our endeavor to create a comprehensive Handwritten Questions Grading System.

Introduction:

The RIMES database, an acronym for "Reconnaissance et Indexation de données Manuscrites et de fac similÉS" (Recognition and Indexation of handwritten documents and faxes), was thoughtfully curated to serve as a crucible for assessing automatic recognition and indexing systems for handwritten letters. This dataset finds particular relevance in scenarios such as handwritten communication sent via mail or fax, connecting individuals with companies or administrations.

Data Collection:

The process of collecting this database involved engaging volunteers who were willing to contribute their handwritten letters in exchange for gift vouchers. To maintain privacy and anonymity, each volunteer assumed a fictitious identity that corresponded in gender to their actual identity. These volunteers were presented with up to 5 distinct scenarios, each drawn from a palette of 9 realistic topics. The scenarios covered a wide spectrum, including changes to personal data, requests for information, customer account management, contract modifications, complaints about service quality, payment issues, reminders, and claims with specific contextual elements. Volunteers were encouraged to craft letters in their own words, with a flexible layout, as long as they used white paper and wrote legibly in black ink.

The success of this campaign was evident in the enthusiastic participation of more than 1,300 individuals who contributed to the creation of the RIMES database. This collective effort has resulted in a database that comprises a staggering 12,723 pages, housing 5,605 unique handwritten letters, each spanning two to three pages.

Accessibility:

The data used in the evaluations of this dataset is also available for download, further enhancing its value in the realm of research and experimentation. Researchers can access the data, particularly in the context of line-level evaluations, through the following link: RIMES-2011-Lines.zip

Database Description: Historical Datasets

In the course of assembling a robust dataset for our Handwritten Questions Grading System project, we have encountered datasets of historical import like Saint Gall Database and

Washington Database. While these may not possess the same quantitative depth as our prior acquisitions, they furnish us with invaluable perspectives on historical handwriting styles, which may exhibit distinctive characteristics such as ornate calligraphy, variations in ink types, and unique penmanship conventions that were prevalent in their respective eras.

Nonetheless, it is essential to be acutely aware that if employed in the phase of handwritten recognition for evaluating written examination papers, these datasets may introduce a potential source of bias. The primary concerns revolve around the marked disparities in vocabulary usage, handwriting conventions, and many other aspects that have evolved significantly over time. As a result, incorporating such historical datasets could lead to suboptimal performance and misinterpretations in the grading process.

To ensure the utmost accuracy and fairness in our Handwritten Questions Grading System, it is prudent to prioritize datasets more closely aligned with the modern educational context, where vocabulary, writing styles, and other linguistic nuances have evolved in tandem with contemporary curricula. By doing so, we can better uphold the integrity of our grading system and minimize potential biases, ultimately providing a more reliable and equitable assessment of students' work.

Saint Gall Database

Data Set:

The Saint Gall database comprises a handwritten historical manuscript with unique characteristics:

- Originating from the 9th century
- In the Latin language
- Authored by a single writer
- Written in Carolingian script
- Utilizing ink on parchment

The original manuscript is preserved at the Abbey Library of Saint Gall in Switzerland. The manuscript images were made accessible online through the e-codices project, and a text edition was attached at the page-level by the Monumenta project. To enhance the dataset, binarized and normalized text line images were contributed. In summary, the dataset includes:

- Page images in JPEG format (300dpi)
- Binarized and normalized text line images
- Text edition at the page-level (please note that word spelling, capitalization, and punctuation may deviate from the image)

Using the semi-automatic process, the following ground truth components were created:

- Text line locations
- Word locations
- Transcription at the line-level (corresponding exactly with the image)

It's essential to note that during text line extraction, only the main text region was covered. Ornamented initial characters and some of the capitalized headings were excluded.

Statistics:

The Saint Gall database consists of:

- 60 pages
- 1,410 text lines
- 11,597 words
- 4,890 word labels
- 5,436 word spellings
- 49 letters

Download:

To access the Saint Gall database, we kindly request that you register before downloading. Once registered, you can obtain the Saint Gall database via the following link: [saintgalldb-v1.0.zip]

The archive includes a README file with comprehensive information about the data formats employed. Furthermore, the training, validation, and test set IDs used in the original publication [1] are provided for automatic transcription alignment. Please be aware that, with regard to this task, word locations were only necessary for the validation and test sets and are not available for the training set.

Washington Database

Data Set:

The Washington database was constructed using the George Washington Papers housed at the Library of Congress. This dataset embodies characteristics from the 18th century and includes:

- Text in the English language
- Contributions from two distinct writers
- Script in longhand
- Utilization of ink on paper

The original manuscript images have been previously utilized in research, including work by Rath and Manmatha. The Washington database encompassed text line and word images, complete with their transcriptions. The dataset includes:

- Binarized and normalized text line images
- Binarized and normalized word images

The ground truth information contains:

- Transcription at the line-level
- Transcription at the word-level

Statistics:

The Washington database comprises:

- 20 pages
- 656 text lines
- 4,894-word instances
- 1,471-word classes
- 82 letters

Download:

As with the Saint Gall database, we kindly request that you register before downloading the Washington database. Once registered, you can access the Washington database through the following link: [washingtondb-v1.0.zip].

The archive includes a README file with detailed information about the data formats used. Additionally, the training, validation, and test set IDs were provided and used as referenced in previous research

In our quest for comprehensive datasets that align with our research objectives, we've encountered several datasets with specific focuses and applications. While these datasets

serve their intended purposes effectively, they may not be ideal for our task due to their inherent characteristics and limitations.

1. The ICDAR Dataset

Introduction:

The ICDAR Dataset is a renowned collection of labeled data primarily used for optical character recognition (OCR) research. While this dataset has significantly contributed to the advancement of OCR technology, it's primarily tailored to character recognition tasks.

Why it may not be suitable for our task:

While the ICDAR Dataset excels in text recognition, it predominantly caters to the domain of OCR, focusing on individual character recognition. Consequently, it may lack extensive annotations for non-textual regions or objects within the images, making it less versatile for our broader handwritten recognition tasks.

2. IIIT 5K-Word Dataset

Introduction:

The IIIT 5K-Word Dataset is a unique collection of cropped word images harvested from Google image searches. With 5,000-word images, it is tailored for word recognition tasks, particularly for large lexicon cropped word recognition.

Why it may not be suitable for our task:

IIIT 5K-Word Dataset is a strong choice for word recognition tasks. However, its specialized focus on individual words and potential limitations in annotating non-textual content may restrict its applicability to broader handwritten recognition tasks. For more extensive handwritten recognition projects that require diverse content analysis and a larger lexicon, it may be necessary to explore other datasets or resources.

3. NIST Database

Introduction:

The NIST Database, provided by the National Institute of Standards and Technology (NIST), comprises a comprehensive collection primarily designed for OCR (Optical Character Recognition) tasks. It serves as a valuable resource for recognizing machine-printed characters as commonly found in printed books and documents.

Why it may not be suitable for our task:

NIST Database is well-suited for OCR tasks, where characters are machine-printed and relatively uniform, making basic recognition methods effective. However, for Handwritten Recognition, which deals with diverse human handwriting styles, more advanced and complex techniques are typically required to achieve accurate results.