

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/220932957>

IAMonDo-database: An online handwritten document database with non-uniform contents

Conference Paper · January 2010

DOI: 10.1145/1815330.1815343 · Source: DBLP

CITATIONS

41

READS

690

3 authors, including:



Emanuel Indermühle

Universität Bern

11 PUBLICATIONS 268 CITATIONS

[SEE PROFILE](#)



Marcus Liwicki

Université de Fribourg

276 PUBLICATIONS 4,869 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Document Classification with Deep Learning [View project](#)



DIVAServices [View project](#)

IAMonDo-database: an Online Handwritten Document Database with Non-uniform Contents

Emanuel Indermühle
Institute of Computer Science
and Applied Mathematics
University of Bern, CH-3012
Bern, Switzerland
eindermu@iam.unibe.ch

Marcus Liwicki
Knowledge Management
Department
German Research Center for
AI (DFKI), Kaiserslautern,
Germany
marcus.liwicki@dfki.de

Horst Bunke
Institute of Computer Science
and Applied Mathematics
University of Bern, CH-3012
Bern, Switzerland
bunke@iam.unibe.ch

ABSTRACT

In this paper we present a new database of online handwritten documents with different contents such as text, drawings, diagrams, formulas, tables, lists, and markings. It was designed to serve as a standard dataset for the development, training, testing and comparison of methods in the field of handwritten document analysis. The database can serve as a basis for layout analysis, and different segmentation and recognition tasks considering online or just offline information. Its size is 1,000 documents produced by approximately 200 writers including a total of 329,849 online strokes. Few constraints were imposed on the writers when creating the documents. Nonetheless, the database has a stable distribution of the different content types. A software tool was developed to allow easy access to the documents which are stored in InkML. In this paper we also present two experiments which show the challenge this database poses. They may figure as references for further research in this area.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous;
D.2.8 [Software Engineering]: Metrics—*complexity measures, performance measures*

General Terms

1. INTRODUCTION

The understanding of entire handwritten documents is a challenging problem which includes a number of difficult tasks. Given a handwritten document, its layout has to be analyzed to isolate different content types in a first step. These different content types can then be directed to specialized systems, including handwriting, symbol, or table recognizers. While a lot of investigation has been directed to the latter three tasks [3, 11, 18], document layout analysis for handwritten documents, just recently gets increasing attention. This is partly due to the rise of novel online hand-

writing recording devices and their use in the daily tasks. Jain et al. were pioneers in this field [8]. Large companies like HP and Microsoft [2] are joining the effort.

The need for a tailored database supporting the development, training, and testing of systems is a central issue in the field of document analysis. Various databases are available and popular in the community. A well known example is CEDAR [7], which is a database of address related text images for the recognition of addresses on letters. The IAMDB [12] is a dataset of handwritten text line images containing sentences of the LOP corpus. These two examples cover offline handwritten text. For online handwriting, widely used datasets are UNIPEN [6], a large repository of online handwritten texts, IRONOFF [17], a database presenting online text which is mapped to the corresponding offline images and IAMonDB [10] which is designed similarly to IAMDB but with online handwritten text. In the research on Japanese and Chinese handwriting recognition online handwritten character databases are often used [13]. All of these databases are focused on handwritten text only. Databases covering entire documents containing drawings, images, and structured text are the UW-I and UW-II databases [15] as well as the more recent ISRI dataset [14] which contain journal articles and memorandums in English and Japanese. However, they are limited to machine printed documents.

In this paper we introduce the IAMonDo-database¹ which is the first database of online handwritten documents with different content types made available to the public. It consists of 1,000 documents containing handwritten text, drawings, diagrams, formulas, tables, lists, and marking elements arranged in an unconstrained way. The database is designed for the development of algorithms for segmentation, content type distinction, and document annotation recognition. Six different classes of content types occur in this database. It is also possible to use the database to distinguish between text and non-text. Moreover, it can be used for recognition and segmentation tasks such as document zone segmentation, formula recognition, symbol recognition, table recognition, text line separation, word segmentation, and handwriting recognition.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DAS '10, June 9-11, 2010, Boston, MA, USA

Copyright 2010 ACM 978-1-60558-773-8/10/06 ...\$10.00

¹IAMonDo-database stands for IAM Online Document Database. IAM is the abbreviation of "Institut für Informatik und angewandte Mathematik"

In Section 2 the design of the IAMonDo-database is described. Its documents are stored in the flexible InkML language (see Section 3), and annotations for document zones down to individual words are available (see Section 4). In this paper we also present some results, demonstrating the challenge this database poses (Section 6). In Section 7 conclusions are drawn and future developments are discussed.

2. DESIGN AND ACQUISITION

The purpose of the database is in the first place to serve as a dataset for the analysis and development of methods distinguishing between different content types in online handwritten documents. In the second place it should assist in research about recognition of document annotation. The two requirements arising therefrom are the need of different content types and the need that the documents should be annotated (we refer to these annotations as *markings* to avoid confusion with the ground truth of the documents which is referred to as *annotations* of the digital ink).

2.1 Database Contents

We assume that the online handwritten documents are typically created in the context of meetings or note taking during lessons. Different content types used in such a context are considered when creating this database; see the following list for details:

Text block: all text which is neither structured as a list nor as a table does belong to this category. Moreover, a text block must not occur as a label in a diagram.

Drawing: graphs, diagrams, maps, symbols or freehand objects.

Diagram: same as drawings, but may include text labels.

Formula: on one or multiple lines.

Table: with and without ruling.

List: containing one word per line.

It can be expected that in real situations most of the content is text. However, to study text vs. non-text distinction, certain amount of drawings and diagrams should be included. This is considered in the IAMonDo-database (See Figure 1). Another element of these documents are *markings* which are applied after the writer has created a document. The marking elements, listed below, are a subset of the Microsoft gesture set (see Figure 2 for an illustration).

1. Underlining of text
2. Marking of text on one or multiple lines by angles on the top left as start mark and angles on the bottom right as end mark
3. Marking of text enclosing it in square bracket.
4. Marking of entire text lines by a vertical stroke on the right, or left side of the text block
5. Encircling

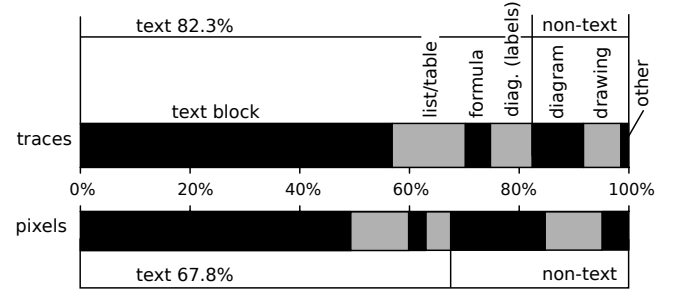


Figure 1: Different content types and their representation in the database. *Other* denotes strokes structuring the document, like ruling in tables, arrows, or separating lines.

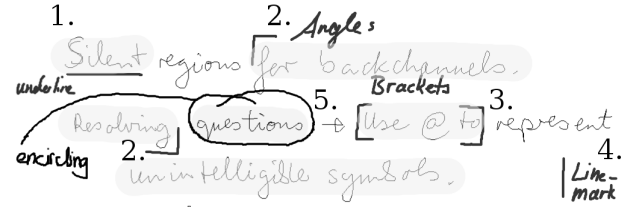


Figure 2: The different marking elements applied to documents.

6. Text labels annotating these markings

7. Strokes connecting a marking with the annotating text label

2.2 Templates

To ensure that each document contains sufficiently large amounts of text, it has been decided to control this by giving templates to the contributors which they had to copy. This also solved the problem of writers not knowing what to write on the given white paper. An additional advantage of using templates is the possibility to control the content distribution.

Having control over then contents leads to the decision to use a language corpus as a text source. In the field of linguistics, several large text collections, known as language corpora, are available. Often, they offer detailed information about the text as for example if a word is a noun, a verb, or another word class. It was decided to use the Brown corpus [5] which is a collection of American English texts covering different fields of writing. For this database the following categories were considered:

- Press: reportage
- Press: editorial
- Press: reviews
- Religion
- Popular lore

- Belles letters, biography, essays
- Miscellaneous: government and house organs
- Education and scientific publications
- Fiction: general
- Fiction: mystery
- Fiction: adventure

Also the other content types were predefined in the templates. Due to legal concerns, drawings, diagrams and formulas have been obtained from Wikipedia² and Wikimedia Commons³. The templates are generated from the following sets.

- Consecutive sentences obtained from the Brown text corpus [5] (category A, B, C, D, F, G, H, J, K, L, and N).
- 200 Drawings obtained from Wikimedia Commons³.
- 200 Diagrams (which are drawings with text labels) obtained from Wikimedia Commons³.
- 200 Formulas obtained from Wikipedia².
- Nouns from the Brown text corpus [5].
- Random numbers.

When generating the templates two conflicting requirements arise. On one hand, the template's content should be randomly compiled so as not to allow any prediction about it. On the other hand, some rules must be applied to guarantee the desired distribution. The following list indicates the rules which are applied during generation of one template.

- Text: A random sentence is selected and the following sentences are appended as long as the text is shorter than 200 characters.
- One random diagram is included.
- With a probability of 0.5 either a random formula or a random drawing is added.
- With a probability of 0.5 either a list or a table is added.
- With equal probability the table is without ruling, with ruling for the title line, with just horizontal ruling, or with a fully ruled grid.
- The text in the table is either aligned left, right, or it is centered.
- The table has 2, 3, or 4 columns and 2 to 6 rows.
- The table contains random nouns in the first column and row, and random numbers of random length in the rest of the cells
- The list has 2 to 7 random nouns

²Wikipedia: <http://www.wikipedia.org>

³Wikimedia Commons: <http://commons.wikimedia.org>

2.3 Acquisition

For the acquisition the Logitech IO2 Pen, powered by technology of Anoto, has been used as a recording device. It has several advantages compared to a tablet PC or an electronic white board. First, it is accurate. Second, besides the coordinates of the digital ink it delivers time and pressure information. Therefore, the writing speed as well as the force at every sampling points can be calculated. Then, the acquisition process is quite easy because no tablet has to be adjusted and no whiteboard must be installed. It needs just a pen and paper to be handed out to the writer, who can do the writing when he or she has time. Afterwards, the pens can be plugged in a cradle and the digital ink is transferred to the host computer.

One of the main problems when acquiring a database is the recruitment of writers. We asked 200 persons to produce 5 documents each, which took about 40 minutes per writer. The writers were mainly students at the institutes of the authors.

During a recording session printed instructions and five templates were given to each writer and he or she was asked to copy each template's content on a sheet of paper. There was no further supervision. The instructions contain a small form to collect meta data of the writer. So the following information can be linked to each individual document.

- A writer id which allows one to say if two documents are from the same writer or not
- The writer's gender
- If the writer is left, or right-handed
- The age
- The native language
- The nationality
- The grade of education
- The profession

In the instructions the writers are requested to copy the content of the template to the paper. They are asked to rearrange the content, i.e. to make it different from the template. The writers were encouraged to use multiple columns for the text, to break apart content elements, to structure the documents with arrows and separating lines, and to simplify or reduce drawings and diagrams. The latter has been added, because the writer tended to spend a lot of time correctly copying the drawing while only a sketchy style was requested. They were asked to correct misspelled words by canceling and not by overwriting it. No further constraints on how to create the document were imposed on them. When they finished the document they were asked to mark (annotate) it. Doing this at the end more likely reflects the workflow in a realistic context. On Figure 3 and 4 one can see how a writer copied the content of a template to the document and then added the markings.

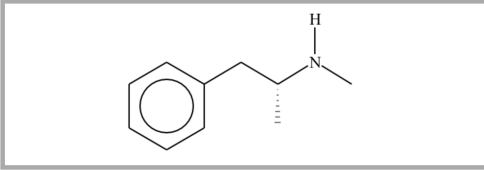
As it was expected, the writers did not always behave according to the instructions given to them. However, the

Vorlage/Template 1

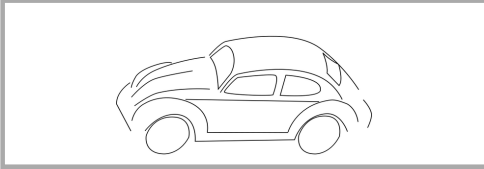
1. Text

And he had a feeling -- thanks to the girl -- that things would get worse before they got better. They had the house cleaned up by noon, and Wilson sent the boy out to the meadow to bring in the horses.

2. Schema/Diagram



3. Zeichnung/Drawing



4. Liste/List

- rancher
- functions
- daughter

Figure 3: Template as it has been presented to the writer.

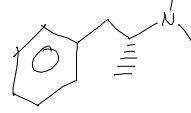
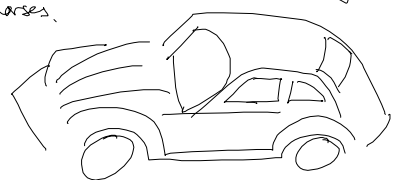
results are very good. Some unexpected behavior reduced the desired diversity of the documents, others increased it. Although they have been asked to rearrange the content, the writer often placed the text on the top of the document. In a majority of cases the use of multiple columns, breaking of content elements and structuring with arrows or separating lines has been omitted. On the other hand about half of the documents are not used in portrait layout but the paper was turned and used in landscape layout which was not expected and resulted in more diversity.

2.4 Additional Information

The database has been split into five disjoint subsets produced by different writers. Each subsets contains approximately 200 documents. For the first four sets (0,1,2,3) two different experimental setups are adopted. In the first one, set 0 and 1 are used for the training, set 2 is used to validate system parameters, and set 3 is the test set. The second setup is a 4-fold cross validation where sets $0 + i$ and $(1 + i \bmod 4)$ are used for training, set $(2 + i \bmod 4)$ for validation, and set $(3 + i \bmod 4)$ for testing, for $i = 0, \dots, 3$. Set 4 is used as an independent test set, which should be used only once for a system.

The database consists of 1,000 documents with 329,849 individual online strokes. The entire database includes 68,441 words, 7,576 text lines, 2,532 table cells, 2,069 list items, and 5,646 diagram labels, which are annotated with transcription. The database contains 905 diagrams, 1456 drawings, 485 formulas, 536 lists, 446 tables, and 1474 text blocks. Of all documents 833 contain markings with 8,260 individual marking elements.

And he had a feeling -- thanks to the girl --
 that things wouldn't get worse before they got
 better. ^{They} wehe
 they took the house
 [cleaned up] by noon,
 and Wilson sent
 the boy out to the meadow to bring in the
 horses.

He stood on the porch
 and watched him struggling
 with the heavy harness,
 and finally went over to help him.

- rancher
- functions
- daughter

/ finally

Figure 4: Document as it has been created by a writer.

Few templates have been copied several times by different writers. This happened accidentally by not keeping track of all templates given to users and by an example which was handed out and has been interpreted as template.

3. FORMAT

To store the document, the InkML language [4], proposed by the World Wide Web Consortium (W3C), has been chosen. It is an XML based language with the full name *Ink Markup Language*. InkML offers high flexibility in how information can be stored and presented. It is also partly used by the International Unipen Foundation⁴ [1].

InkML is intended to preserve the original form of the data, while offering the possibility to create different views of the ink. This is done by two constructs. The first one is the **TraceView** element. This XML tag can either contain other **TraceView** elements or they are referencing to ink data. This allows hierarchical structuring of the data, which is used in the documents of this database. The second construct allows to transform the data by different means. In this work it is used to correct the orientation of the documents by applying an affine transformation to the original coordinate system.

Additionally, InkML allows to include annotations as name-value pairs virtually everywhere. Since this gives a high degree of freedom it is important to define application specific rules. The definition of such rules, however, can not be done with InkML and is therefore specific to this database. We have defined the structure of the annotation in a separated XML file.

⁴International Unipen Foundation: <http://unipen.org/>

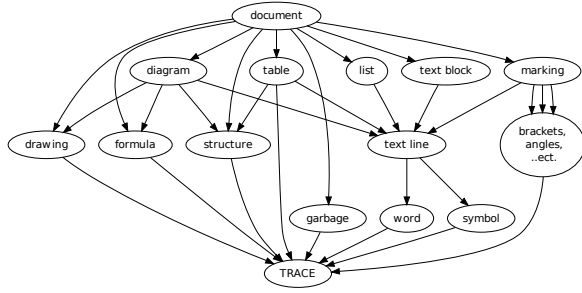


Figure 5: The hierarchical structure of the annotation.

4. ANNOTATION

The database has detailed annotations to generate ground truth in various formats. On the top-level different document zones are labeled with their content type. These zones are divided further into sub-elements, such as text lines, table cells, or drawings which are annotated with their type and transcription if applicable. Text lines or table cells are again divided into individual words which are also annotated with their description. The grammar of the hierarchy represented by a graph is shown in Figure 5.

The hierarchically structured annotation is on the InkML level realized by annotating **TraceView** elements which are structured as desired. Formulas have not been annotated in detail. This can be done in the future, however.

The annotation has been done using the software presented in Section 5. The process of annotating starts with the operator selecting traces which, grouped together, form for instance a word or a drawing. Such trace groups are annotated by one of the content types. These content types are proposed by the software following the defined rules in the additional XML file. Also the transcription, if applicable, is specified. Having a couple of words or other annotated trace groups, further groups can be built constituting text lines, diagrams or other high-level structures. According to the entities selected for grouping, the software proposes appropriate types to annotate this group.

Sometimes, it is hard to recognize the transcription of a given word even for the operator. In these cases it is handy having a template to look at. The whole database has been annotated this way in about 200 hours.

Different forms of ground truth can be generated from the annotation of a document. Some exported formats have already proven useful in our experiments. If document analysis is made considering only offline information, the document can be converted to a color coded image as it is proposed in [16]. Several different codes can be used to color different document zones, text lines, or even individual words. This is useful for page, line, or word segmentation tasks as well as for content distinction. For handwriting recognition it is possible to export individual text line or word images, accompanied by the transcription. For online handwritten

document analysis, feature vectors of traces or even individual sampling points can be exported directly, labeled by a class number representing the intended classification.

5. SOFTWARE

InkAnno is the name of the software tool used to handle the documents of this database. It is written in the Java programming language and therefore it runs on a number of platforms. To export images it depends on the JAI⁵ (Java Advanced Imaging) library and for PDF export the iText⁶ library is required to be installed.

The core of the software is a library which reflects the structure of an InkML document in a changeable model. Using this library the software offers a graphical user interface which can be used to display the documents, edit the **TraceView** tree, modify annotations, and export the document into different formats. These formats include images, color-coded images, PDFs, lists of feature vectors, and InkML files of course. New exporters can be developed with little effort.

As InkML is quite complex its not recommended to reimplement interpreters but to use existing libraries. As of now the library proposed here is the only one implementing enough language elements to interpret this database. However, there are other projects, like the InkML toolkit project⁷ heading towards full InkML compatibility.

InkAnno is published under the open source GPL license and therefore freely accessible⁸.

6. EXPERIMENTS

The purpose of this database is primarily to develop and evaluate methods for the distinction of different content types. In this section we present two methods which try to solve the text vs. non-text distinction problem.

6.1 Trace classification

The first methods was introduced by Jain et al. [8]. It is a classification of the individual ink traces taking two simple features into account: the length of the trace and the accumulated curvature. Jain et al. tested this method on a dataset which is not available to the public. It was not clear from their paper what classifier was used, but it was probably linear regression. They achieved a recognition rate of 99.1% on a test set with 35,882 text traces and 930 non-text traces. In our experiment rerunning the method on the database presented in this paper, an SVM classifier is used instead. SVMs, similarly to linear regression, are capable of finding a linear separation between two classes. Taking advantage of the kernel trick, however, it is possible to increase the dimensionality of the feature space. The hyperplane found by the SVM which is linear in the higher dimensional space corresponds to a nonlinear separation when projected to the original space. Using this classifier we can expect to get a result at least as good as with a linear regression.

⁵Java advanced imaging library: <https://jai.dev.java.net/>

⁶iText library: <http://www.lowagie.com/iText/>

⁷See: <http://inkmltk.sourceforge.net>

⁸Visit the homepage of our research group for more information <http://www.iam.unibe.ch/fki>

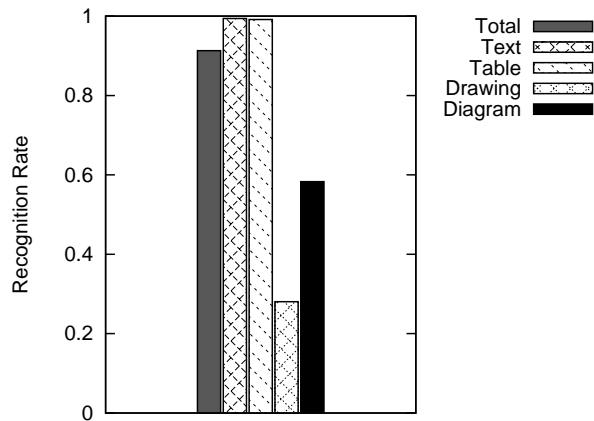


Figure 6: Results of text vs. non-text trace classification, displayed for individual content types.

As one can see in Figure 6 the recognition rate of the SVM on this database is 91.3%, which is significantly lower than the 99.1% reported in [8]. The difference can be explained by the fact that the dataset used by Jain et al. is highly unbalanced. Even if every trace would be classified as text, the recognition rate would be 97.5%. With a 4:1 text ratio this is not the case with the database presented here. Since the features are the same and the classifiers can be seen as performing equally well the difference must be in the challenge posed by the two datasets. The IAMonDo-dataset with its balanced content and varying style poses a similar problem as real world documents and therefore serves as a suitable base to measure performance of text versus non text distinction.

6.2 Connected component classification

The second method considers only offline information. After applying a connected component analysis (8-side neighborhood) on the document, 82 features are extracted from the individual connected components.

Keyesers et al. [9] have investigated different feature sets for document zone classification. They proposed a well performing set of run-length histograms and connected component statistics which can be extracted very fast. The features used in the current paper for the classification of the connected components are similar to those. The feature set consists of run-length histograms of black and white pixels along the horizontal, vertical, and the two diagonal directions. Each histogram uses eight bins, counting runs of length $\leq 1, 3, 7, 15, 31, 63, 127$, and ≥ 128 . For each histogram the mean and the variance are considered. Adding the width and height of the considered connected component it sums up to a total of 82 features.

As in the trace classification an SVM is trained on a training set and achieves 94.4% (see Figure 7) of correctly classified

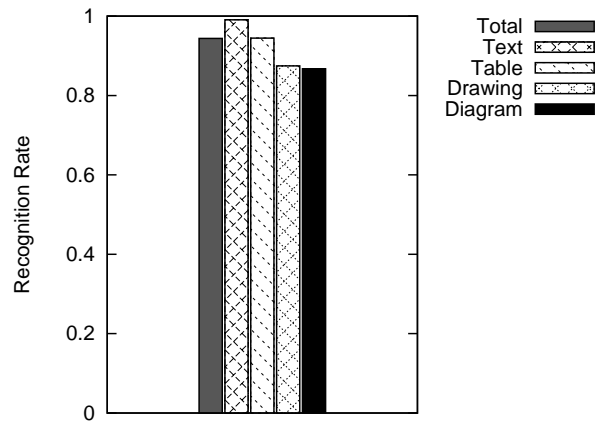


Figure 7: Results of text vs. non-text connected component classification, displayed for individual content types.

pixels⁹. Note that in this experiment only offline information is used.

7. CONCLUSION AND OUTLOOK

A database of online handwritten documents containing text blocks, drawings, diagrams, formulas, lists, and tables as well as markings has been described in this paper. The documents have been written in an unconstrained manner and reflect documents generated in a realistic context. With 200 writers many different styles in writing and drawing can be found. However, the local and global distribution of the content types is kept rather stable over all individual documents. The database can serve as a basis for content distinction, marking recognition, layout analysis, as well as formula recognition, handwriting recognition, word and text line segmentation, and many more tasks. Various formats of ground truth can be generated thanks to the detailed annotation of the documents. A software system has been presented which allows easy access to the documents. It can also be used to display, modify and transform the documents.

Two text vs. non-text distinction methods have been presented in this paper. The first one is based on a method proposed in [8] which is applied to show the challenge this dataset poses when compared to another dataset. The other experiment applies simple methods using just offline information of the database. The results may serve as a benchmark for future systems to be developed for this database.

At our institute we plan to use the database for ongoing research. The database itself is considered complete at the current stage. Some errors of the annotation which may occasionally be detected will be corrected. The software tool, InkAnno, will undergo further development, hopefully in collaboration with other institutes.

⁹Please consider that the percent of correctly classified pixel in the connected component classification system can not directly be compared to the percent of correctly classified traces as it is used for the trace classification.

Acknowledgments. We would like to thank all the individuals who contributed to the database described in this paper. Particular thanks go to Karin Indermühle for her recruiting skills and Amrei Schröttke for annotating the whole database. Further thanks are due to IM2.VP for financial support.

8. REFERENCES

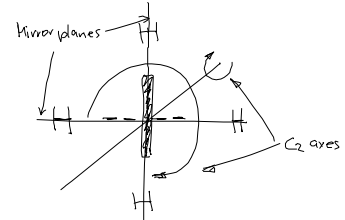
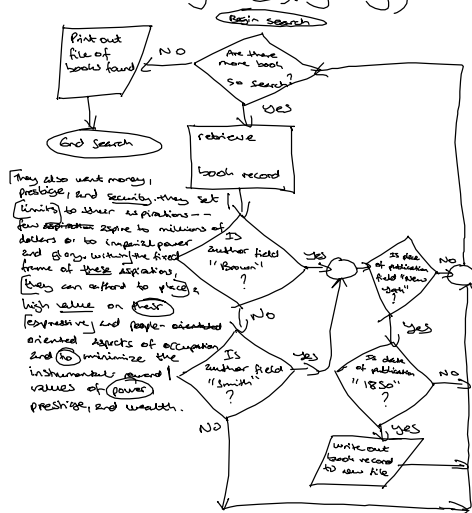
- [1] M. Agrawal, K. Bali, and S. Madhvanath. Upx: A new xml representation for annotated datasets of online handwriting data. In *Proc. 8th Int. Conf. on Document Analysis and Recognition*, pages 1161–1165, 2005.
- [2] C. M. Bishop, M. Svensen, and G. E. Hinton. Distinguishing text from graphics in on-line handwritten ink. In *Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition*, pages 142–147, Washington, DC, USA, 2004. IEEE Computer Society.
- [3] H. Bunke. Recognition of cursive Roman handwriting—past, present and future. In *Proc. 7th Int. Conf. on Document Analysis and Recognition, Edinburgh*, pages 448–459. IEEE, 2003.
- [4] Y.-M. Chee, M. Froumentin, and S. Watt, editors. *Ink markup language (InkML)*. World Wide Web Consortium, 2006.
<http://www.w3.org/TR/2006/WD-InkML-20061023>.
- [5] W. Francis and H. Kucera. *Manual of information to accompany a standard sample of present-day edited American English for use with digital computers*. Department of Linguistics, Brown University, 1979.
- [6] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet. UNIPEN project of on-line data exchange and recognizer benchmarks. In *Proc. 12th Int. Conf. on Pattern Recognition*, volume 2, pages 29–33 vol.2, Oct 1994.
- [7] J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.
- [8] A. K. Jain, A. M. Namboodiri, and J. Subrahmonia. Structure in on-line documents. In *Proc. 6th Int. Conf. on Document Analysis and Recognition*, pages 844–848, 2001.
- [9] D. Keysers, F. Shafait, and T. M. Breuel. Document image zone classification – a simple high-performance approach. In *Proc. 2nd Int. Conf. on Computer Vision Theory and Applications*, pages 44–51, 2007.
- [10] M. Liwicki and H. Bunke. IAM-OnDB – an on-line English sentence database acquired from handwritten text on a whiteboard. In *Proc. 8th Int. Conf. on Document Analysis and Recognition*, volume 2, pages 956–961, 2005.
- [11] J. Lladós, E. Valveny, G. Sánchez, and E. Martí. Symbol recognition: Current advances and perspectives. In *Proc. 4th Int. Workshop on Graphics Recognition Algorithms and Applications*, pages 104–127, London, UK, 2001. Springer-Verlag.
- [12] U.-V. Marti and H. Bunke. The IAM-database: an English sentence database for offline handwriting recognition. *Int. Journal on Document Analysis and Recognition*, 5:39–46, 2002.
- [13] M. Nakagawa and M. Onuma. On-line handwritten japanese text recognition free from constrains on line direction and character orientation. In *Proc. 7th Int. Conf. on Document Analysis and Recognition*, pages 519–523, Edinburgh, Scotland, 2003.
- [14] T. A. Nartker, S. V. Rice, and S. E. Lumos. Software tools and test data for research and testing of page-reading ocr systems. In *International Symposium on Electronic Imaging Science and Technology*, volume 1, pages 37–47. SPIE, 2005.
- [15] I. Phillips, J. Ha, R. Haralick, and D. Dori. The implementation methodology for a cd-rom english document database. In *Proc. 2nd Int. Conf. on Document Analysis and Recognition*, pages 484–487, Oct 1993.
- [16] F. Shafait, D. Keysers, and T. M. Breuel. Pixel-accurate representation and evaluation of page segmentation in document images. In *Proc. 18th Int. Conf. on Pattern Recognition*, volume 1, pages 872–875, 2006.
- [17] C. Viard-Gaudin, P. M. Lallican, P. Binter, and S. Knerr. The ireste on/off (ironoff) dual handwriting database. In *Proc. 5th Int. Conf. on Document Analysis and Recognition*, pages 455–458, Los Alamitos, CA, USA, 1999. IEEE Computer Society.
- [18] R. Zanibbi, D. Blostein, and R. Cordy. A survey of table recognition: Models, observations, transformations, and inferences. *Int. Journal of Document Analysis and Recognition*, 7(1):1–16, 2004.

APPENDIX

A. SAMPLE DOCUMENTS

breakdown cases run	
town	2264 74
artists	65 1336

$$= 5x(2y^2 + 3xy - y)$$



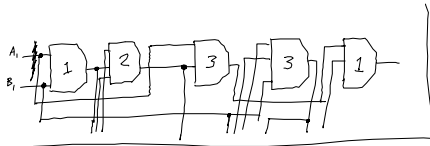
There are many causes [for this change]. One of the most important is economic. Business leaders are aware now that they

$$D(k|n, n) \approx \phi\left(\frac{k-n_0}{\sqrt{2k \ln n}}\right) - \phi\left(\frac{k-1-n_0}{\sqrt{2k \ln n}}\right)$$

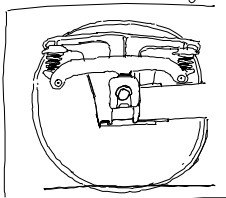
suffer greatly from any outbreak of violence. They are putting strong pressure on their police departments to keep order.

Even

- clarification
- result
- type
- chance
- disaster
- hope
- wife

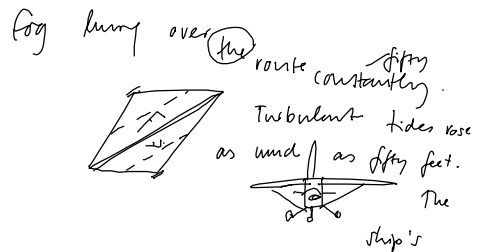


After inspiring this, I think we should certainly follow through on it, he declared. It has become our responsibility and I hope that the [Citizens Group] will spearhead the movement. He said he would not be surprised if some of the more than 30 members of the group



- time
- arrival
- worry

one interested in moving on the required non-partisan ballot for posts on the charter commission.



Compass was useless because of the nearness of the magnetic North Pole.

As the berg grew larger, Hinds

was forced to turn assumption brittle voyage home South into what pays 6'367 866 198 now intatement 15 60 70 Ungava Bay, an inlet of the Great Strait.