The HHD Dataset

1st Irina Rabaev
Software Engineering Department
Shamoon College of Engineering
Be'er Sheva, Israel
irinar@ac.sce.ac.il

2nd Berat Kurar Barakat
Computer Science Department
Ben-Gurion University of the Negev
Be'er Sheva, Israel
berat@post.bgu.ac.il

3rd Alexander Churkin
Software Engineering Department
Shamoon College of Engineering
Be'er Sheva, Israel
alexach3@sce.ac.il

4thJihad El-Sana
Computer Science Department
Ben-Gurion University of the Negev
Be'er Sheva, Israel
el-sana@cs.bgu.ac.il

***Abstract*—Benchmark datasets are important in document image processing field, as they allow to analyze different approaches and compare their performances in a fair manner. There exist benchmark datasets for several alphabets such as Latin, Arabic and Chinese, but not the Hebrew alphabet. In this paper, a handwritten Hebrew dataset, HHD, is introduced. The HHD dataset is collected from hand-filled forms, and accompanied by their ground truth at character, word and text line levels. Presently, the dataset contains around 1000 document images, and we continue to further enlarge it. To the best of our knowledge, this is the first comprehensive corpus of Hebrew handwritten documents, and we believe it will help leveraging Hebrew documents processing and document processing in general. The dataset can be useful for various research applications, such as word spotting, word recognition, text line alignment, and writer identification. The initial small subset of the HDD for character classification can be downloaded from *https://www.cs.bgu.ac.il/~berat/data/hhd_dataset.zip* together with the training and test sets subdivisions. We also provide baseline results for character classification on this initial subset. In the near future, the full HHD dataset will be made freely available to the research community.**

***Keywords*-Handwritten document image dataset, Hebrew handwritten documents, Ground truth**

I. Introduction

Benchmark datasets are of tremendous importance, as they provide a platform for evaluation and fair comparison of different methods. Over the past decades, several datasets supporting the evaluation of word spotting, OCR, text line extraction, and writer identification have been introduced [1]–[7]. However, there is a lack of Hebrew handwritten dataset for developing, benchmarking and improving methods that are frontiers in the Hebrew handwritten document processing. Without a doubt, having such a standard dataset will help leveraging document image processing in Hebrew. The lack of a Hebrew dataset is even more relevant for historical document processing. Annotating historical documents is a tedious and expensive task and often an expert is required to read ancient texts. To overcome the limitation of small training set, some learning algorithms can utilize contemporary document images for the training process, requiring only a small set of annotated historical documents to adjust final parameters ([8]–[10]).

Considering the issues mentioned above, we present a Hebrew Handwritten Dataset (HHD). The HHD dataset contains cursive Hebrew script written by different writers from different backgrounds and age groups, both by native and non-native Hebrew speakers. It is composed of around 1000 scanned images of handwritten forms and their ground truth at the character, word, and text line levels. We also describe the method used to annotate the scanned forms. The initial small subset of the HDD, which consists of images of isolated characters, is available for downloading[1] together with the subdivision into training and test sets. We also provide baseline results for character classification on this initial subset. In the near future, the complete HHD will be made publicly available for initiating researches in handwritten Hebrew related problems, such as word spotting and recognition, text line alignment, and writer identification. Besides, HHD can be used to prove the robustness of handwritten image processing methods in general.

II. Review of the existing datasets

The only Hebrew dataset we are aware of is the Pinkas historical dataset [11], which contains 30 pages from the Pinkas manuscript, together with its ground truth at page, line and word levels. We are not aware of any modern comprehensive Hebrew dataset, or any other Hebrew dataset. However, there is a rich variety of publicly available handwritten datasets in other languages.

The most popular historical datasets are the DIVA-HisDB [12], the historical IAM [13]–[15], and the IMPACT datasets, which consist of manuscript images in Latin languages. Among the modern widely used datasets is the IAM dataset [1]. The IAM is an English sentence dataset for handwriting recognition at the line and word levels. It includes 1066 handwritten forms written by 400 different

[1] https://www.cs.bgu.ac.il/~berat/data/hhd_dataset.zip

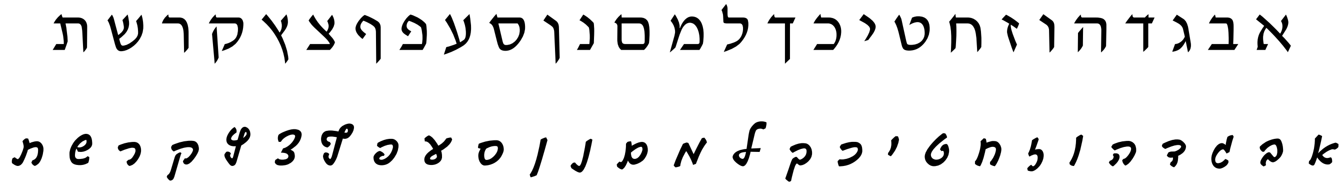


Figure 1. Hebrew alphabet: printed (top row) and cursive (bottom row).

writers. A relatively small but widely used handwritten dataset is MNIST [2], comprising of 10 classes of handwritten digits images. Subsequently, the EMNIST [3] dataset has been introduced, which involves numerical digits and both uppercase and lowercase letters, and constitutes a larger and more challenging dataset.

Among the most popular Arabic datasets is the IFN/ENIT dataset [4]. It contains Tunisian town names written by 411 writers. A more comprehensive and rich vocabulary Arabic dataset is the KHATT dataset [5]. It consists of 1000 handwritten forms written by 1000 writers. These forms include 2000 randomly selected paragraphs and cover the various shapes of Arabic characters. The AHTID/MW dataset [6] contains the text lines and word images written by 53 native Arabic speakers, and is used for Arabic recognition, word segmentation, and writer identification tasks.

From the above literature review, it is clear that the comprehensive large dataset of handwritten document images is a crucial resource for Hebrew image processing research.

III. Description of Hebrew Script

Hebrew is written from right to left using the Hebrew alphabet, whose letters are not similar to any other alphabet. Hebrew alphabet is a set of 22 consonant-only letters, five of them have additional form when used at the end of the word. In handwriting, cursive Hebrew letters are used, whereas, in a printed text, block Hebrew letters are used. Cursive Hebrew letters are more circular and considerably vary from their equivalent Hebrew block letters. This reality is the major reason that reveals the demand for a handwritten Hebrew dataset. Letters have no case and no vowels. Sometime diacritics are placed above and below letters to specify the pronunciation. However, most texts appear without the diacritics, and the pronunciation is implied by the word and the context. Figure 1 and Figure 2 illustrate the above descriptions of Hebrew letters. Hebrew is characterized by high similarities among letters. Figure 3 illustrates this similarities; each row shows the group of three very similar letters. This property of the script makes Hebrew document image processing even more challenging.

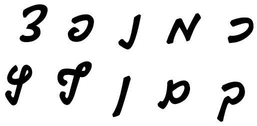


Figure 2. Five letters that have a different final form used at the end of words; the final forms are displayed beneath the regular forms.

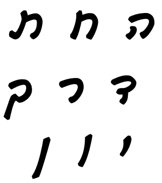


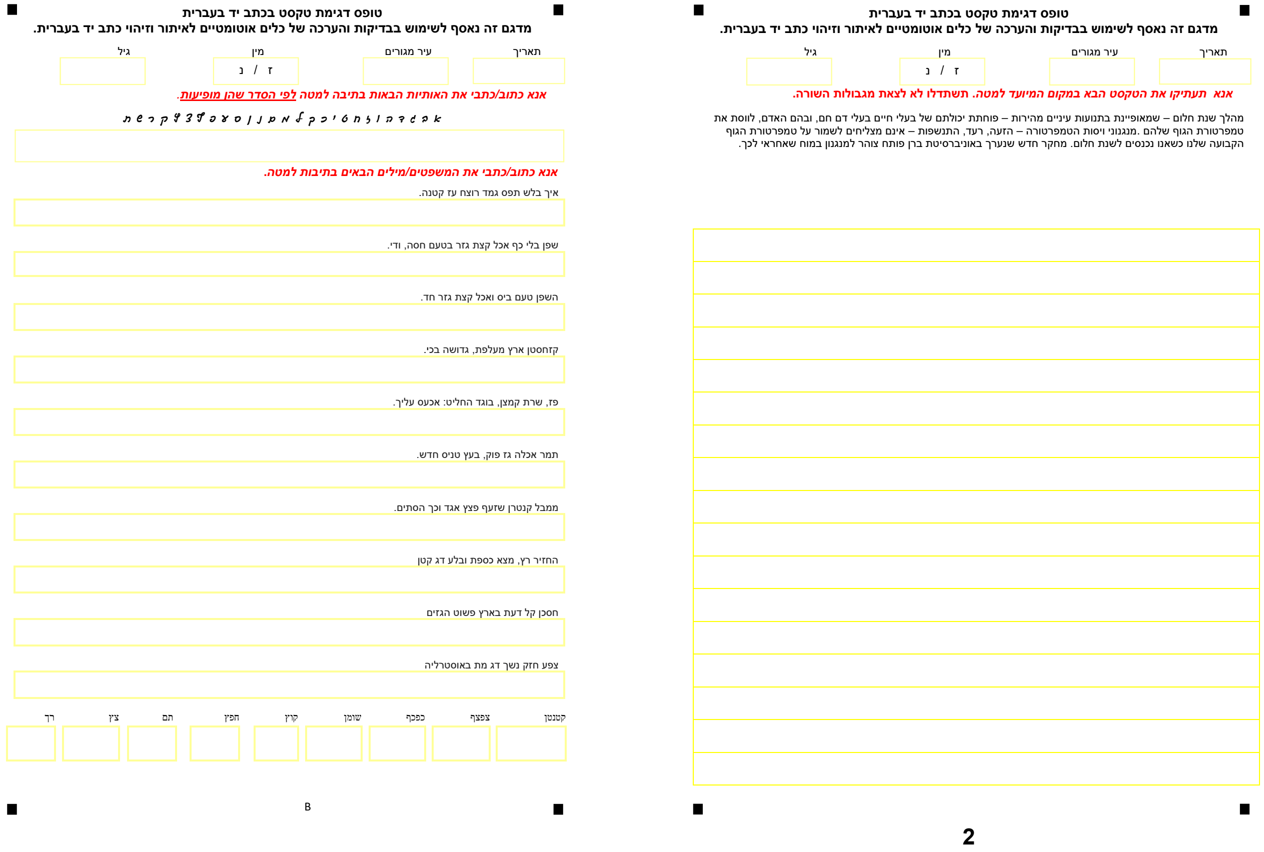
Figure 3. Hebrew is characterized by high similarities among letters. Each row shows the group of three similar letters.

IV. Description of the HHD Dataset

The HHD dataset contains around 1000 handwritten forms written by different writers. These forms were scanned in color with the resolution of 600 dpi, and are stored in TIFF format. The design of the forms was carefully planned. There are 63 variations of the forms. The first 13 forms are marked by the letters $A - M$. They contain text fields for sentences and isolated words. The next 50 forms are marked by numbers $1 - 50$. Each of them contains a text paragraph from one of the four categories: (1) general news, (2) scientific articles, (3) children books, and (4) economy news. Altogether, the forms contain approximately 2500 different classes of words. The scanning process may introduce document skew. To make future document processing easier, four small black squares were added at the corners of each form (see Figure 4). In addition, yellow colour was used to mark the boundaries of each text field (since it is relatively simple to eliminate yellow channel).

For data collection, the forms were given to individuals from different age groups and educational backgrounds, both native and non-native Hebrew speakers. The participants used a pen or a pencil, and no restrictions (such as the pen colour or its type) were imposed. Each participant filled between 1 to 10 different forms depending on their willingness. Therefore, the dataset can be also used in writer identification researches.

Each of the forms $A - M$ is divided into several text fields, 22 fields in forms $A - E$, and 14 fields in forms $F - M$. The printed instructions above each text field entry describe how to fill the field. The first row of text fields comprises statistical information, such as date, town, sex and age. We supposed such information might be of interest for future researches, however, these fields were filled voluntarily. The



טופס דגימת טקסט בכתב יד בעברית

מדגם זה נאסף לשימוש בבדיקות והערכה של כלים אוטומטיים לאיתור וזיהוי כתב יד בעברית.

תאריך

עיר מגורים

מין ז / נ

גיל

אנא כתוב/כתבי את האותיות הבאות בתיבה למטה <u>לפי הסדר שהן מופיעות</u>.

א ב ג ד ה ו ז ח ט י כ ך ל מ ם נ ן ס ע פ ף צ ץ ק ר ש ת

אנא כתוב/כתבי את המשפטים/מילים הבאים בתיבות למטה.

איך בלש תפס גמד רוצח עז קטנה.

שפן בלי כף אכל קצת גזר בטעם חסה, ודי.

השפן טעם ביס ואכל קצת גזר חד.

קזחסטן ארץ מעלפת, גדושה בכי.

פז, שרת קמצן, בוגד החליט: אכעס עליך.

תמר אכלה גז פוק, בעץ טניס חדש.

ממבל קנטרן שזעף פצץ אגד וכך הסתים.

החזיר רץ, מצא כספת ובלע דג קטן

חסכן קל דעת בארץ פשוט הגזים

צפע חזק נשך דג מת באוסטרליה

קטנטן צפצף כפכף שומן קוץ חפץ תם ציץ רך

B

טופס דגימת טקסט בכתב יד בעברית

מדגם זה נאסף לשימוש בבדיקות והערכה של כלים אוטומטיים לאיתור וזיהוי כתב יד בעברית.

תאריך

עיר מגורים

מין ז / נ

גיל

אנא תעתיקו את הטקסט הבא במקום המיועד למטה. תשתדלו לא לצאת מגבולות השורה.

מהלך שנת חלום – שמאופיינת בתנועות עיניים מהירות – פוחתת יכולתם של בעלי חיים בעלי דם חם, ובהם האדם, לווסת את
טמפרטורת הגוף שלהם .מנגנוני ויסות הטמפרטורה – הזעה, רעד, התנשפות – אינם מצליחים לשמור על טמפרטורת הגוף
הקבועה שלנו כשאנו נכנסים לשנת חלום. מחקר חדש שנערך באוניברסיטת ברן פותח צוהר למנגנון במוח שאחראי לכך.

2

Figure 4. Examples of different forms from the HHD dataset: sentences based on pangrams (on the left), and a text paragraph (on the right). The text boxes in the first row of each form are provided to fill the personal information (voluntarily). The black rectangles at the corners of the page are used for aligning the scanned image.

next row contains the text field with isolated letters of the Hebrew alphabet (including final forms of five letters). The participants were asked to write each letter exactly once in the order of their appearance in the form. For the rest of the rows, the writers were asked to copy the text printed above each text field. The final row of forms $A-E$ contains nine text boxes of isolated words; forms $F-M$ contain text-line rows only. The text the participants were asked to write was carefully chosen. One of our goals was to generate a balanced dataset on character level, i.e. the dataset will contain approximately the same amount of each character of the alphabet. This is a challenging task, as there are characters that are used frequently and characters that are comparably rarely used. This feature is present in all languages and is not specific for Hebrew. To accomplish the stated goal, the sentences in forms $A-E$ were chosen to be *pangrams*. A pangram is a sentence where every character of a given alphabet appears at least once. For example, ”The quick brown fox jumps over the lazy dog” is the famous English pangram. In almost all pangrams on forms $A-E$, each character appears *exactly* once. On the contrary, the sentences on forms $F-M$ were taken from famous children’s stories and are not balanced on the character level. On the other hand, they are meaningful for writers, and thus are easier to copy. Table I lists the frequency of each alphabet letter in the forms $A-M$. As can be seen, the character distribution in forms $A-E$ (pangrams) is balanced, even for the five final letters (coloured in grey). The five final letters rarely appear in forms $F-M$ (children’s stories). On the other hand, forms $F-M$ linguistically constitute more ”natural” sentences. Figure 4 (left) illustrates form B, and Table II summarises the total numbers of characters, words, and sentences present in each form.

Each of the form $1-50$ contains a text paragraph. Four different categories of texts were chosen: general news, scientific articles, children’s books, and economy news. On average, each form includes 3.86 sentences, 61.5 words, and 290 letters. Figure 4 (right) illustrates one of the forms, and Table III presents the total number of text paragraphs, sentences and words per each category in forms $1-50$.

	Pangrams					Children Stories							
Form Letter	A	B	C	D	E	F	G	H	I	J	K	L	M
א	10	11	9	10	9	7	16	22	17	12	12	22	14
ב	10	11	10	10	11	18	20	21	19	17	12	15	15
ג	10	10	9	11	10	10	1	4	9	4	3	6	4
ד	10	10	9	12	9	5	14	8	13	16	8	13	13
ה	10	10	10	10	10	34	21	40	24	23	35	28	26
ו	13	11	10	11	17	26	23	45	28	47	52	31	33
ז	10	10	9	10	10	1	1	9	3	2	2	2	1
ח	11	10	10	10	12	7	5	11	7	13	16	6	8
ט	12	12	11	12	10	1	4	3	2	6	7	1	6
י	11	12	12	11	13	31	42	49	34	38	34	28	21
כ	10	11	10	9	11	12	8	8	10	9	11	5	3
ך	9	6	3	7	4	1	2	1	4	5	3	4	6
ל	10	12	10	11	11	15	20	28	27	32	31	18	30
מ	9	10	9	9	10	23	16	17	18	23	26	14	12
ם	9	7	5	7	5	8	17	25	14	15	17	7	7
נ	10	7	6	6	7	11	15	18	9	7	11	7	7
ן	10	8	6	7	6	2	1	0	3	3	4	2	3
ס	10	10	9	10	10	2	5	2	2	3	4	3	3
ע	10	11	9	10	10	7	12	11	8	16	6	8	11
פ	11	13	9	12	12	3	5	6	7	3	5	6	9
ף	10	4	6	5	5	0	0	0	0	1	2	1	2
צ	12	10	11	10	8	2	1	1	2	4	3	4	7
ץ	10	8	3	7	8	0	0	0	0	2	0	3	1
ק	12	12	12	12	12	13	5	15	4	6	5	4	7
ר	11	12	10	11	11	16	14	29	14	17	18	11	18
ש	11	9	9	9	10	16	13	11	10	16	17	11	11

Table I
Distribution of the characters in each form. The final letters are emphasized in gray. It can be noted that the letters distribution is balanced for forms $A - E$, even for the five final letters. The five final letters appear rarely in forms $D - M$; these forms linguistically constitute more ”natural” sentences.

	Pangrams					Children’s stories							
Form Text Object	A	B	C	D	E	F	G	H	I	J	K	L	M
Letters	281	268	237	259	262	289	296	401	298	357	354	274	287
Words	82	77	73	76	75	69	66	91	82	83	84	64	76
Sentences	10	10	10	10	10	11	11	11	11	11	11	11	11

Table II
The total number of letters, words and sentences in forms $A - M$.

Category	Number of paragraphs	Total number of sentences	Total number of words
General news	10	40	653
Scientific articles	20	63	1248
Children book	10	59	625
Economy news	10	31	549

Table III
The total number of text paragraphs, sentences and words per each category in forms $1 - 50$.

V. Annotation of the HHD Dataset

The generated ground truth (GT) of each text line, word and character is represented by its bounding rectangle and the corresponding transcription, and is saved in PAGE [16] XML file format. Figure 5 illustrates samples of the annotated forms with the superimposed ground truth at character, word, and text line levels. The personal information of the participants was blacked out. Figure 6 illustrates the handwritten form with superimposed transcription above the text line (top row), words (middle row), and characters (bottom row).

The structure of the forms facilitates automatic ground truth generation. First, the skew introduced by the scanning process is corrected. Each document image is aligned horizontally using the black squares in the corners of the document (see Figure 4). Next, the handwritten text lines are extracted using the coordinates of the corresponding text boxes. The most challenging part is to locate and annotate each word and character in the text. For this task, we experimented with two approaches. The first one is to utilize projection profiles for locating words inside text lines, and then use the connected

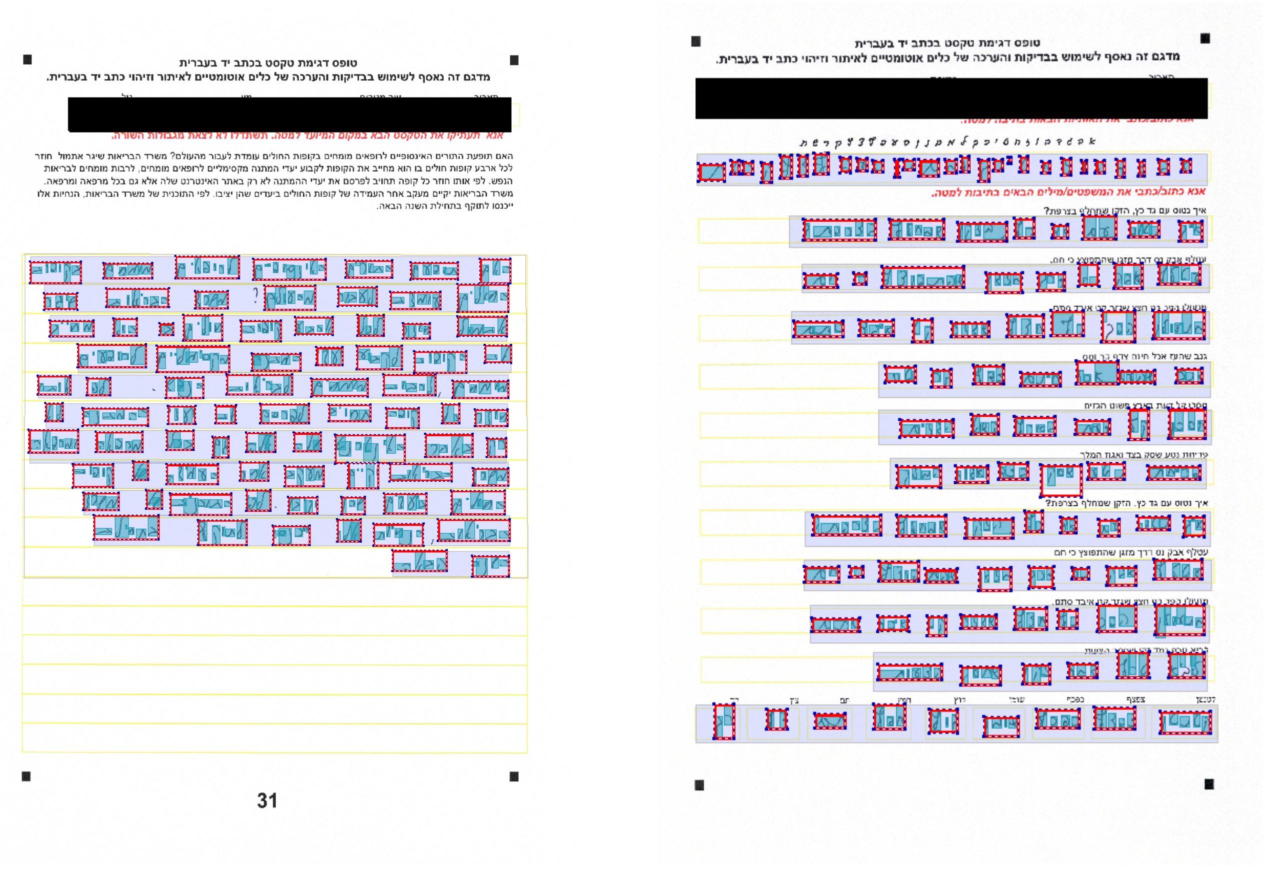

טופס דגימת טקסט בכתב יד בעברית
מדגם זה נאסף לשימוש בבדיקות והערכה של כלים אוטומטיים לאיתור וזיהוי כתב יד בעברית.
[redacted]
31
טופס דגימת טקסט בכתב יד בעברית
מדגם זה נאסף לשימוש בבדיקות והערכה של כלים אוטומטיים לאיתור וזיהוי כתב יד בעברית.
[redacted]

Figure 5. Annotated handwritten forms with superimposed ground truth, showing bounding rectangles of text lines, words and characters. The personal information of the participants was blacked out.

components of each word to extract characters. The second approach utilizes the alignment algorithm described in [17]. This algorithm allows locating and annotating words and characters simultaneously. We used HOG descriptors together with χ^2 distance for the similarity measure of the images inside the alignment algorithm. After comparing the results and the running time of these two approaches, we saw that the projection profiles based approach provides satisfactory results and its running time is much faster than the alignment based approach. So we decided to adopt the projection profiles based approach for words and character extraction. At the final stage, the automatic annotations are verified and corrected (if needed) by a human.

VI. Initial Experiments

The manual verification process is very slow and still is in progress. Meantime, we experimented with a small subset of the dataset, which we call HDD_v0. HDD_v0[1]consists of images of isolated Hebrew characters together with training and test sets subdivision. We have experimented with three different Neural Networks to set the baselines for characters classification: simple CNN with three hidden layers, AlexNet [18] and ResNet [19]. Each model was trained for 100 epochs. The results are summarized in Table IV. We can see that ResNet achieved the best results. We have performed an error analysis, and found that most of the errors are between classes of characters that are very similar, especially the three

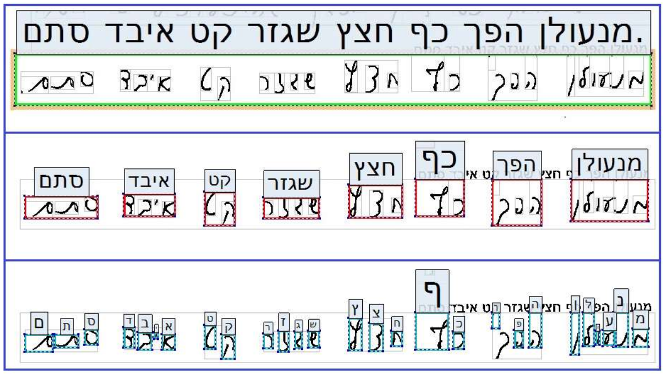


Figure 6. Annotated handwritten forms with superimposed transcription above the text line (top row), words (middle row) and characters (bottom rows).

	Train accuracy	Test accuracy
Simple CNN	96.62	72.57
AlexNet	99.55	78.21
ResNet	100	84.9

Table IV
CHARACTER CLASSIFICATION RESULTS ON HDD_v0 (SMALL SUBSET OF THE HDD)

characters in the last row in Figure 3. The only difference between them is their length. The error analysis confirms the challenging nature of the Hebrew script, and shows that there is a large room for improvement.

VII. CONCLUSION AND FUTURE WORK

The lack of a standard dataset for Hebrew document images motivated us to develop the HHD - a handwritten dataset of Hebrew documents images. The dataset is composed of scanned images of handwritten forms and their ground truth at text line, word and character levels. The first 13 forms contain isolated sentences and words. Five of these forms are based on pangrams sentences, and thus represent a completely balanced set of characters images. Another 50 forms contain text paragraphs from four categories: general news, scientific articles, children's books, and economy news. The structure of the forms utilizes automatic ground truth generation. Presently, the HDD contains around 1000 document images, and we continue to extend it. The dataset can serve as a basis for research in Hebrew handwritten document images analysis, including both segmentation-based and segmentation free word spotting, word recognition, text alignment, and writer identification. In addition, the HHD dataset can be used for the initial training of learning-based algorithms with limited training data, for example, in the case of historical documents. Learning algorithms usually require a lot of training data that is not always available for historical documents, thus, only a small set of historical document will be needed to tune the final parameters.

The manual verification process is slow, and still is in progress. Meantime, we performed initial experiments for character classification on a small subset of HDD. The obtained results show that there is a large room for improvement. In the future research, we are planning to include an additional ground truth format for character recognition, which will be easy to use with machine learning methods (such as MNIST format). We are also planning to run baseline experiments for word spotting and word recognition on full HHD, and publish further baseline results.

To the best of our knowledge, HHD is the first dataset of modern document images in handwritten Hebrew, which provides a diversity of writing styles and fully labeled at character, word, and text line levels. The HHD contributes to the heterogeneity of benchmarking standards and we are going to make it publicly available.

ACKNOWLEDGMENT

Authors would like to thank Coral Burg, Emil Sandler and Mark Rabayev for their part in implementing the software for automatic ground truth extraction from the handwritten forms.

REFERENCES

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

[2] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: an extension of MNIST to handwritten letters," *arXiv preprint arXiv:1702.05373*, 2017.

[4] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, H. Amiri *et al.*, "IFN/ENIT-database of handwritten Arabic words," in *Proc. of CIFED*, vol. 2. Citeseer, 2002, pp. 127–136.

[5] S. A. Mahmoud, I. Ahmad, M. Alshayeb, W. G. Al-Khatib, M. T. Parvez, G. A. Fink, V. Märgner, and H. El Abed, "KHATT: Arabic offline handwritten text database," in *2012 International Conference on Frontiers in Handwriting Recognition*. IEEE, 2012, pp. 449–454.

[6] A. Mezghani, S. Kanoun, M. Khemakhem, and H. El Abed, "A database for Arabic handwritten text image recognition and writer identification," in *2012 international conference on frontiers in handwriting recognition*. IEEE, 2012, pp. 399–402.

[7] S.-H. Cha and S. N. Srihari, "Assessing the authorship confidence of handwritten items," in *Proceedings Fifth IEEE Workshop on Applications of Computer Vision*. IEEE, 2000, pp. 42–47.

[8] S. Sudholt and G. A. Fink, "PHOCNet: A deep convolutional neural network for word spotting in handwritten documents," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 277–282.

[9] G. Retsinas, G. Sfikas, and B. Gatos, "Transferable deep features for keyword spotting," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 2, no. 2, 2018, p. 89.

[10] P. Krishnan and C. Jawahar, "HWNet v2: An efficient word image representation for handwritten documents," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 22, no. 4, pp. 387–405, 2019.

[11] B. K. Barakat, I. Rabaev, and J. El-Sana, "The Pinkas dataset," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 732–737.

[12] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, and R. Ingold, "DIVA-HisDB: A precisely annotated large dataset of challenging medieval manuscripts," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 471–476.

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

[14] A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stolz, "Automatic transcription of handwritten medieval documents," in *2009 15th International Conference on Virtual Systems and Multimedia*. IEEE, 2009, pp. 137–142.

[15] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription alignment of Latin manuscripts using hidden Markov models," in *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*. ACM, 2011, pp. 29–36.

[16] S. Pletschacher and A. Antonacopoulos, "The PAGE (page analysis and ground-truth elements) format framework," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 257–260.

[17] I. Rabaev, R. Cohen, J. El-Sana, and K. Kedem, "Aligning transcript of historical documents using dynamic programming," in *Document Recognition and Retrieval XXII*, vol. 9402. International Society for Optics and Photonics, 2015, p. 94020I.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.