CVL-Database: An Off-line Database for Writer Retrieval, Writer Identification and Word Spotting

Florian Kleber, Stefan Fiel, Markus Diem and Robert Sablatnig
Computer Vision Lab
Institute of Computer Aided Automation
Vienna University of Technology
Favoritenstraße 9/1832, 1040 Vienna
Email: dir@caa.tuwien.ac.at

***Abstract*—In this paper a public database for writer retrieval, writer identification and word spotting is presented. The CVL-Database consists of 7 different handwritten texts (1 German and 6 English Texts) and 311 different writers. For each text an RGB color image (300 dpi) comprising the handwritten text and the printed text sample are available as well as a cropped version (only handwritten). A unique ID identifies the writer, whereas the bounding boxes for each single word are stored in an XML file. An evaluation of the best algorithms of the ICDAR and ICHFR writer identification contest has been performed on the CVL-database.**

I. Introduction

Document analysis methods dealing with the recognition of handwritten data need training samples and datasets for evaluation and objective comparison, since the "*collection of the data and the production of the corresponding transcriptions (labeling) are expensive and time consuming tasks it is desired to reuse existing databases*" [1]. The International Association for Pattern Recognition (IAPR), Technical Committee 11 collects databases together with the Ground Truth (GT) data and specification for Online and Offline Handwriting Recognition, Signature Verification, Scene Text Recognition and e.g. machine-printed documents for layout analysis (http://www.iapr-tc11.org/). The archiving of public datasets will allow an easy access and the reproduction of published results.

Marti and Bunke [2] summarized available databases for offline recognition (CEDAR [3], NIST [4], CENPARMI [5], Rimes [6]). Due to the restrictions of handwriting databases available (isolated characters, single words, limited vocabulary) Marti and Bunke collected and published the IAM-database [2]. The IAM database 3.0 (updated) contains full English sentences, 115320 words and is produced by 657 writers. Due to the large vocabulary the database is proposed for handwriting recognition.

The Qatar University Writer Identification dataset (QUWI) consisting of Arabic and English texts has been presented by Somaya Al Maadeed et al. [7]. Each writer has written two English and two Arabic texts (one random and one specified text for each language) and provided additional information: gender, hand used, year of birth interval, and nationality. The random texts can be used for text independent writer identification. The entire database consists of 1017 different writers and will be made public in the future.

The CVL-database has 311 writers and was designed for writer retrieval and identification. The database consists of 7 (27 writers) respectively 5 (284 writers) different texts (101069 words at all). The text statistics (number of words, number of unique words, lexical diversity) is presented in Section II. Additionally each page is labeled and provides the coordinates of the bounding boxes of each word (punctuations are not annotated) encoded using an XML-file. Thus, the CVL-database can also be used for the evaluation of word-spotting methods. In contrast to the IAM database the number of pages of each writer is distributed more equally (see Section II). State-of-the-Art methods of writer identification (submitted to the ICDAR 2011 [8] and ICFHR 2012 competitions [9]) have been evaluated in Section III.

The CVL-database can be downloaded publicly (no registration required) at http://caa.tuwien.ac.at/cvl/research/cvl-database/. Additionally an XML Parser (C++) and a GT Viewer is provided.

This paper is organized as follows: Section II describes the design of the CVL-database and the automated labeling. Further, the cross-evaluation, the texts used and the word statistics are presented. In Section III the results of State-of-the-Art writer identification methods are presented. Finally, a conclusion is given in Section IV.

II. CVL-Database

The CVL-database consists of images with cursively handwritten German and English texts which have been choosen from literary works. Samples of the following texts have been used:

- Text 1 Edwin A. Abbot - Flatland: A Romance of Many Dimension (90 words).
- Text 2 William Shakespeare - Mac Beth (47 words).
- Text 3 Wikipedia - Mailüfterl (74 words, under CC Attribution-ShareALike License).

Text 4 Charles Darwin - Origin of Species (52 words).
Text 5 Johann Wolfgang von Goethe - Faust. Eine Tragödie (50 words).
Text 6 Oscar Wilde - The Picture of Dorian Gray (65 words).
Text 7 Edgar Allan Poe - The Fall of the House of Usher (73 words).

Figure 1 shows an example of a filled-out form. All pages have a unique writer id and the text number (separated by a dash) at the upper right corner, followed by the printed sample text. The text is placed between two horizontal separators. Individuals have been asked to write the text beneath the printed text using a ruled undersheet to prevent curled text lines. The layout follows the style of the IAM database [2].

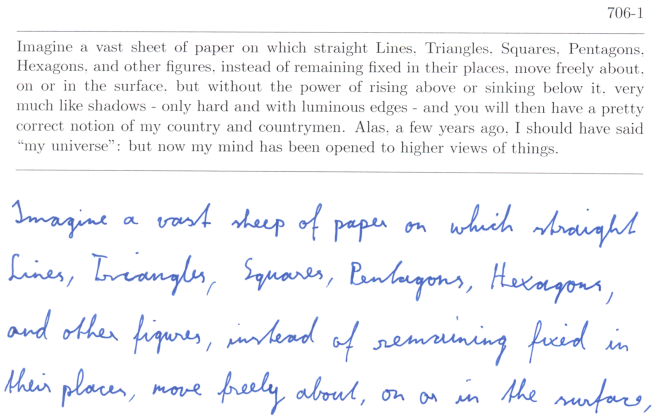

Figure 1. Detail of an example page of the CVL Dataset (Writer ID 706, Text #1).

The filled forms have been scanned with a Lexmark X652de scanner at a resolution of 300 dpi and a color depth of 24 bit. All postprocessing steps of the scanner (e.g. sharpening, color dropout) have been turned off, and all images are stored as LZW encoded TIFFs. Figure 2 shows the distribution of the number of text pages regarding different writers of the CVL and the IAM database. It can be seen that the CVL-database is almost equally distributed (all writers have written text 1-5). In contrast, for the IAM database appr. 350 writers have written just one page (different texts), while one wrote 60 pages. The advantage of equally distributing the texts over all writers is the fact, that the effect of the writing style on algorithms evaluated gets minimized.

Table I shows statistics for all texts. For each text the number of words, the number of unique words and the type-token ratio is given. The type-token ratio TTR (lexical diversity measure) [10] is defined as:

$$TTR = \frac{total\ unique\ words}{total\ word\ count} \quad (1)$$

Table I
CVL DATASET - TEXT STATISTICS.

text	total word count	total unique words	TTR (1) [%]
E. Abbot #1	90	73	81
W. Shakespeare #2	47	41	87
Wikipedia #3	74	56	75
C. Darwin #4	52	41	78
J. Goethe #5	50	39	78
$\sum$1-5	313	216	69
O. Wilde #6	65	44	67
E. Poe #7	73	53	72
$\sum$ 1-7	451	292	64

For word spotting methods all pages have been labeled. Figure 3 shows a labeled example page. It can be seen that all words of the text are surrounded by a bounding box (labeled punctuations are left blank), which has been automatically calculated (see Section II-A) and manually checked by 2 students (see Section II-B).

```
<?xml version="1.0" encoding="UTF-16" ?>
<PcGts>
    <Metadata>
        <Creator>Vienna UT</Creator>
    </Metadata>
    <Page imageFilename="0901-6.tif"
          imageHeight="3507" imageWidth="2480">
        <dkTextLines cropH="2518" cropW="2518"
             cropX="0" cropY="918">
            <textlines xE="338.346" xS="290.115"
                 yE="360.331" yS="369.977" />
            ...
        </dkTextLines>
        <AttrRegion area="428373" attrType="4"
             fontAngleRad="0.350019"
             fontSize="41.7165" fontType="2"
             medianWordHeight="857.5">
            <minAreaRect>
                <Point x="404" y="2273" />
                <Point x="404" y="401" />
                <Point x="2197" y="401" />
                <Point x="2197" y="2273" />
            </minAreaRect>
            <AttrRegion area="305160" attrType="3"
                 ...>
                <AttrRegion area="11341" attrType="1"
                     fontType="2" text="Verweile">
                ....
                </AttrRegion>
            </AttrRegion>
        </AttrRegion>
        ...
    </Page>
</PcGts>
```

Listing 1. Sample GT XML

The information is stored in an XML file. An example file is shown in Listing 1. The root node *Page* contains the file-name and the image size, which is followed by the *dkTextLines* node. This node defines the position of the crop area (region of the handwritten text) as well as all *textlines* (division line between two text lines). Subsequently to *textlines*, text areas are defined by the nodes *AttrRegion*. Each *AttrRegion* has an attribute *attrType* which defines the hierarchy: value 4 defines the global region (entire page), value 3 defines a text block, value 2 defines a textline and

value 1 defines a single word. Additional attributes are the area of the bounding box, the font type (1=machine printed, 2=handwritten), the median word height (for text lines), and the defined text.

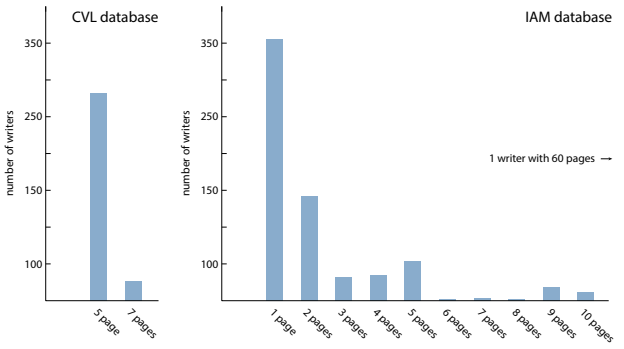

Figure 2. Distribution of the number of pages written by different writers of the CVL (left) and IAM (right) database.

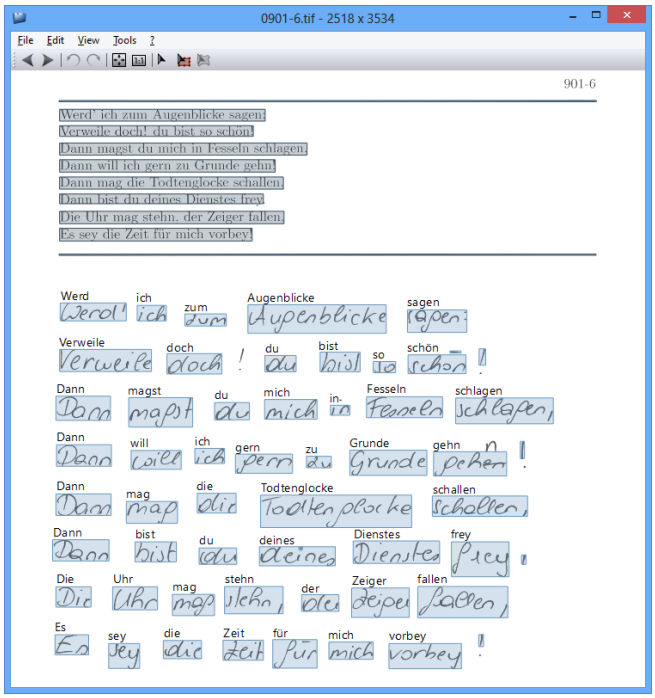

Figure 3. Labeled example page of the CVL Dataset (Text #5). Note that punctuations are partially marked with bounding boxes. Since they have no groundtruth annotated, they are ignored during the processing.

In Listing 1 the *AttrRegion* with *fontType* 2 (line 19) defines the attributes of the word "Verweile". The coordinates (basic points) of the bounding box are defined by the attributes of *Point* of the node *minAreaRect*.

In Section II-A the pre-processing steps and layout analysis methods for the labeling of the CVL-database are explained.

A. Automated GT Labeling

As a pre-processing step the skew of the scanned pages is corrected. The method is based on the text's gradients in combination with a Focused Nearest Neighbor Clustering (FNNC) of interest points [11]. The combination of both methods (gradient and FNNC) can be used for slanted handwritten text, and using integral images [12] allows a fast implementation. The method is described in detail in [11]. In order to localize and classify text regions, words are estimated by means of Local Projection Profiles (LPP). Then, automatically detected text lines split word blobs which are falsely merged between text lines. Finally, minimum area rectangles are found by means of Rotating Calipers. The text clustering aims at grouping the previously classified word blobs. Therefore, words are clustered according to text lines and paragraphs. The text annotation per rectangle was carried out automatically based on the GT text.

For the GT stored in the XML file the minimum area rectangle is replaced by the bounding box of the word blobs. A description of the layout analysis is presented in [13].

B. Cross-Evaluation of the CVL-Database GT and Statistics

Automated labeling results have been evaluated using a visualization tool (see Figure 3) by 2 students independently. They corrected errors made by writers like missing, misspelled or hyphenated words as well as errors caused by the automated annotation. Therefore, the visualization tool allows for manual insertion/deletion and editing of GT rectangles and for deleting/inserting the annotated text. The word order or spelling of the GT text cannot be edited in order to guarantee a consistent tagging throughout the database. Hence, if a writer missed a word or swapped some words, the GT rectangles and the annotation of these words were deleted. The students corrected 74300 (74.4%) word rectangles from the automated tagging. Though this is a high percentage of corrections, the annotation was semi-automated. Hence, they had to delete one noisy rectangle in order to correct all subsequent labels.

In order to minimize the error of the manual correction, a cross-validation of both operators was carried out (see Figure 4). Therefore, a tool was developed that compares the consistency of both annotations for all writers. The tool displays all pages where the rectangle of at least one word does not overlap for more than 40% (or vice versa) to its correspondent. Having detected these errors, they were displayed to two different operators that could either tag one of the annotations as correct or manually open the current page and correct the annotation. The cross-validation was carried out two times by two operators respectively. In the first cross-validation stage, a total of 2208 rectangles (2.21%) were inconsistent and corrected. This corresponds to 333 incorrect pages out of 1600. The second cross-validation detected 191 (0.2%) wrong rectangles.

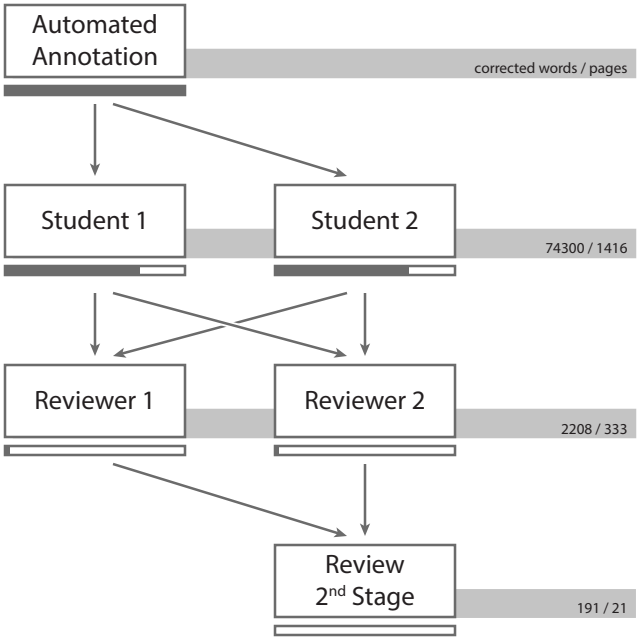

Figure 4. GT tagging methodology. There were a total of 99904 words and 1600 pages.

As stated before, the database should consist of 101069 words. However, since some writers missed words – or in rare cases whole sentences/pages – we end up with a total of 99904 different words. Hence, 1165 words are missing.

III. Evaluation of Writer Identification Methods on the CVL-Database

For the evaluation of Writer Identification Methods on the CVL-Database we contacted the participants of the "ICDAR 2011 Writer Identification Contest" [8] and the "ICFHR2012 Competition on Writer Identification" [9]. Following authors provided us their binaries for evaluation:

- CS-UMD [14] - Rajiv Jain and David Doermann, University of Maryland, College Park, USA
- QUQA A [15] - Abdelaali Hassaine, Somaya Ali S. Al-Ma'adeed, and Ahmed Bouridane, Computer Science and Engineering Department, Qatar University, Doha, Qatar
- QUQA B [15] - Abdelaali Hassaine, Somaya Ali S. Al-Ma'adeed, and Ahmed Bouridane, Computer Science and Engineering Department, Qatar University, Doha, Qatar
- TEBESSA A [16] - Chawki Djeddi, Computer Science Department, Cheikh Larbi Tebessi University, Algeria and Labiba Souici-Meslati, Computer Science Department, Badji Moktar University, Algeria
- TEBESSA B [16] - Chawki Djeddi, Computer Science Department, Cheikh Larbi Tebessi University, Algeria and Labiba Souici-Meslati, Computer Science Department, Badji Moktar University, Algeria
- TEBESSA C [17] - Chawki Djeddi, Computer Science Department, Cheikh Larbi Tebessi University, Algeria, Imran Siddiqi, Department of GS & AS, Bahria University, Pakistan, Labiba Souici-Meslati, Computer Science Department, Badji Moktar University, Algeria, Abdellatif Ennaji, LITIS Laboratory, Rouen University, France
- TSINGHUA [18] - Xiaoqing Ding, Department of Electronic Engineering, Tsinghua University, Beijing, China

To obtain an equally distributed dataset only the pages 1-5 of all writers were used. The evaluation was done using the soft *TOP-N* and the hard *TOP-N* criterion which were already used in the ICDAR 2011 and ICFHR 2012 competitions. For every document image of the database the distance to all other document images is calculated and this distance is sorted from the most similar to the least similar document image. For the soft *TOP-N* criterion at least one document in the *N* most similar has to be from the same writer and the percentage of all documents with a correct hit is calculated. For the hard *TOP-N* criterion it is considered as a correct hit if all document images in the *N* most similar images are from the same writer. The values of *N* used for the soft criterion are 1, 2, 5 and 10 as they are already used in the ICDAR 2011 and ICFHR 2012 contest. For the hard criterion we use 2,3 and 4. Since we have 5 document images per writer, 4 is the maximum value of *N* for this criterion.

Additionally an evaluation for writer retrieval was carried out. Writer retrieval is the task of finding all documents in the database which have been written by the same writer as a reference document. For this task, again, all distances from the reference document image to all other document images are calculated. We examine the first *N* document images and calculate how many of them are from the same writer as the reference document. Using all document images in the database once, we calculate the percentage of how many document images have been found correctly. For this evaluation we use also 2,3 and 4 as values for *N*.

The evaluation for the soft criterion on all methods is presented in Table II. For the hard criterion the results are listed in Table III. The evaluation for writer retrieval is presented in Table IV.

IV. Conclusion

A freely available database for writer retrieval, writer identification and word spotting has been presented. The writer is identified by a unique ID, whereas the coordinates of the words bounding box (transcription) are stored in an XML file. The CVL-database consists of 311 writers and 5-7 different texts which are appr. equally distributed. State-of-the-Art writer identification algorithm have been

Table II
SOFT EVALUATION USING THE ENTIRE DATASET (%)

Method	Top 1	Top 2	Top 5	Top 10
CS-UMD	97.9	98.4	99.1	99.4
QUQA A	30.5	41.4	57.5	67.1
QUQA B	92.9	96.0	97.9	98.3
TEBESSA A	69.8	80.4	89.5	94.4
TEBESSA B	96.0	97.0	97.8	98.1
TEBESSA C	97.6	97.9	98.3	98.5
TSINGHUA	97.7	98.3	99.0	99.1

Table III
HARD EVALUATION USING THE ENTIRE DATASET (%)

Method	Top 2	Top 3	Top 4
CS-UMD	90.0	71.0	48.3
QUQA A	5.7	0.5	0.1
QUQA B	84.9	71.5	50.6
TEBESSA A	44.5	27.4	12.3
TEBESSA B	91.4	83.0	64.6
TEBESSA C	94.3	88.2	73.9
TSINGHUA	95.3	94.5	73.0

Table IV
RETRIEVAL EVALUATION USING THE ENTIRE DATASET (%)

Method	Top 2	Top 3	Top 4
CS-UMD	94.2	87.9	80.6
QUQA A	23.5	19.5	16.7
QUQA B	90.5	86.5	80.4
TEBESSA A	62.5	56.8	50.6
TEBESSA B	94.2	91.5	86.1
TEBESSA C	96.1	94.2	90.0
TSINGHUA	96.8	94.5	90.2

evaluated on the database. The CVL-database can be downloaded at http://caa.tuwien.ac.at/cvl/research/cvl-database/. Additionally an XML-Parser (C++) and a GT-Viewer is provided.

ACKNOWLEDGMENT

We would like to thank all authors who participated in the ICFHR and ICDAR writer identification contest and provided the executable of their method for evaluation purposes. Additionally we would like to thank all individuals who have contributed to the CVL-database, in particular by Janina Ninetta Bratu and Alin Dumitru.

REFERENCES

[1] M. Zimmermann and H. Bunke, "Automatic segmentation of the IAM off-line database for handwritten English text," in *16th International Conference on Pattern Recognition, 2002. Proceedings*, vol. 4, 2002, pp. 35 – 39 vol.4.

[2] 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.

[3] J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550 –554, may 1994.

[4] R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Greecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, and C. Wilson, *The first census optical character recognition systems conference #NISTR 4912*. The U.S. Bureau of Census and the National Institute of Standards and Technology, 1992.

[5] C. Suen, C. Nadal, R. Legault, T. Mai, and L. Lam, "Computer recognition of unconstrained handwritten numerals," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1162 –1180, jul 1992.

[6] F. Slimane, S. Kanoun, H. E. Abed, A. M. Alimi, R. Ingold, and J. Hennebert, "ICDAR 2011 - Arabic Recognition Competition: Multi-font Multi-size Digitally Represented Text," in *Proceedings of the 11th International Conference on Document Analysis and Recognition*, 2011, pp. 1449–1453.

[7] Somaya Al Máadeed, Wael Ayouby, Abdelaali Hassaïne, and Jihad Mohamad Aljaam, "QUWI: An Arabic and English Handwriting Dataset for Offline Writer Identification," in *International Conference on Frontiers in Handwriting Recognition*, 2012, pp. 746–751.

[8] G. Louloudis, N. Stamatopoulos, and B. Gatos, "ICDAR 2011 Writer Identification Contest," in *International Conference on Document Analysis and Recognition*, 2011, pp. 1475–1479.

[9] G. Louloudis, B.Gatos, and N. Stamatopoulos, "ICFHR2012 Competition on Writer Identification, Challenge 1: Latin/-Greek Documents," in *2012 International Conference on Frontiers in Handwriting Recognition*, 2012, pp. 825–830.

[10] Victoria Johansson, "Lexical diversity and lexical density in speech and writing: a developmental perspective," in *Lund Working Papers in Linguistics*, vol. 53, 2008, pp. 61–79.

[11] Markus Diem, Florian Kleber, and Robert Sablatnig, "Skew Estimation of Sparsely Inscribed Document Fragments," in *Proc. of 10th IAPR International Workshop on Document Analysis Systems*, Goldcoast, Australia, 2012, pp. 292–296.

[12] Mohamed E. Hussein, Fatih Porikli, and Larry S. Davis, "Kernel integral images: A framework for fast non-uniform filtering," in *International Conference on Computer Vision and Pattern Recognition, CVPR*, 2008, pp. 1–8.

[13] M. Diem, F. Kleber, and R. Sablatnig, "Text Classification and Document Layout Analysis of Paper Fragments," in *Proceedings of the 11th International Conference on Document Analysis and Recognition*, 2011, pp. 854–858.

[14] Rajiv Jain and David Doermann, "Offline Writer Identification using K-Adacent Segments," in *International Conference on Document Analysis and Recognition*, 2011, pp. 769–773.

[15] A. Hassaiine, S. Al Maadeed, and A. Bouridane, "A Set of Geometrical Features for Writer Identification," in *Neural Information Processing*, ser. Lecture Notes in Computer Science, T. Huang, Z. Zeng, C. Li, and C. Leung, Eds. Springer Berlin / Heidelberg, 2012, vol. 7667, pp. 584–591.

[16] C. Djeddi and L. Souici Meslati, "A texture based approach for Arabic Writer Identification and Verification," in *IEEE International Conference on Machine and Web Intelligence*, Alger, Algerie, 2012, pp. 115–120.

[17] C. Djeddi, I. Siddiqi, L. Souici Meslati, and A. Ennaji, "Multi Script Writer Identification Optimized With Retrieval Mechanism," in *International Conference on Frontiers Handwritting Recognition*, Bari, Italy, 2012, pp. 507–512.

[18] Xin Li and Xiaoqing Ding, "Writer Identification of Chinese Handwriting Using Grid Microstructure Feature," in *Proceedings of the Third International Conference on Advances in Biometrics*, 2009, pp. 1230–1239.