Recognition of Humboldt's Handwriting in Complex Surroundings

Karl-Heinz Steinke, Martin Gehrke, Robert Dzido
University of Applied Sciences and Arts, Hanover; Germany
karl-heinz.steinke@fh-hannover.de

Abstract

In the Botanic Museum of Berlin in Germany exist about 3.5 million dried plants on paper sheets. They were collected in the last two hundred years by many collectors among those was the famous researcher "Alexander von Humboldt". because e every collector left his handwriting on the sheet the question came up if it is possible to find out automatically which sheet belongs to Humboldt's collection. To solve the problem many challenging sub problems have to be solved before. First text regions have to be localized and extracted. Then printing is to be distinguished from handwriting. For the latter an off-line writer recognition procedure had to be developed in order to determine the writer of the handwriting. Often two or more writers can be found on one sheet. To make sure that the suspect writer is correct an interactive method is applied which transforms the static character into a dynamic form. Different mathematical procedures are used such as the reproduction of the write line of individual characters by Legendre polynomials. All methods were proved on the international IAM-database [3]. 93 writers with at least 5 samples were chosen from the IAM-database. By combining multiple characters (up to thirteen), the recognition rate rises considerably and reaches 98.7%. A global statistical approach using the whole handwritten text results in a similar recognition rate of more than 98%. By combining the methods by a ranking method a recognition rate of 99.5% is achieved.

1. Introduction

In the past 200 years many natural scientists collected plants and glued them upon paper sheets. Usually the description of the plant was done by handwriting on the sheets. There exists a big collection in the Botanic Museum of Berlin in Germany. Among the approximately 3.5 million herbarium sheets there are many whose collector was the famous scientist "Alexander von Humboldt". These sheets should be allocated by the analysis of the handwriting contained in it. The problem to find Humboldt's collections automatically among all the sheets involves several challenging milestones which have to be solved. First text regions (printing and handwriting) have to be localized in a complex environment with objects like roots, leaves, stamps, bar codes, yardsticks, color charts etc.. In literature there are approaches to locate texts in pictures as e.g. on book envelopes, cheque forms, in colored announcements, video pictures, internet pictures or general color images. [14] - [17]. In most cases it concerns thereby printed letters which are to be recognized. With the available sheets however the mixture of handwriting of different writers, block letters in different sizes and forms, marked text and bar code information represents a special challenge.

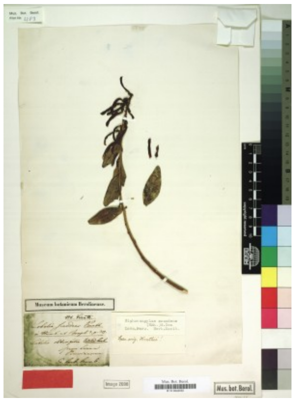


Figure 1: Sheet with Humboldt's writing

After text regions are localized and extracted with the method described in section 2, stamps, barcode, printing and different handwritings have to be distinguished. This is done by texture analysis combined with writer recognition methods. In literature there are many approaches of writer recognition. In our case we are dealing with old static handwritings of writers who are not present. This so-called off-line writer recognition represents a more complicated problem than on-line recognition since no coordinate sequences are available. Research in automatic

978-0-7695-4221-8/10 $26.00 © 2010 IEEE
DOI 10.1109/ICFHR.2010.91

identification of writers focuses mainly on the statistical approach. This leads to the extraction of characteristics such as run lengths [10] and inclination distributions as well as entropy characteristics. Newer approaches, e.g. that of Siddiqi [11] try to combine global and local features with good results. Niels [5] uses character prototypes and differentiates writers on the basis of how often the prototypes occur in a long text. Srihari [6] developed individuality-characteristics for static pictures by extraction of macro and micro features. It was shown that individual characters possess different capabilities of discriminating between writers. Marti [4] analyzes the difference in handwritings by structural characteristics of each text line. Schomaker [8] uses the contour of connected components. Bensefia [1] uses local characteristics which originate from the analysis of the upper contour's minima.

After handwriting is separated from printing and others finally the writer has to be identified. This is done by a combination of a text independent global method and a local character method. The global method uses run length statistics in eight directions. The characters method transfers static handwriting into dynamic coordinate sequences. By the subsequent treatment of these x and y-coordinates with different algorithms, characteristics are obtained from single characters which suggest the possible writer of the characters.

2. Localization of Text regions

In order to reduce the computing time the original color image (see fig. 1) with 600 dpi is transformed to a grey image with 300 dpi. To get vertical lines the Sobel operator is applied by folding the picture with a 3*3 window. By the Sobel filtering connected handwriting and block letters are divided into small almost vertical writing segments which show the character inclination with their inclination. After binarization black and white changes are computed at each position of the image (see fig. 2a). One can recognize that the plant and the color chart are suppressed to a large extent and mainly writing ranges, stamps and the bar code remains. The write lines become now apparent as bright stripes in the image. So there is a high probability for writing to lie in bright stripes of the "black and white changes" image. Writing has to fulfill some more features. 1. Between the write lines the high horizontal contrast drops fast. 2. The writing objects are to possess a defined height (character height) in order to be considered as writing. 3. In a close environment of the objects it is to look writing-like i.e. the texture has certain statistic characteristics (see section 3). 4.Writing objects must have neighbors in the same line. If not all described writing attributes are fulfilled or writing objects are lying outside of bright stripes (see fig. 2a) they are rejected. By a cluster analysis method now neighboring objects are merged to write lines. Boundary boxes (see fig. 2b) are determined by the maximum and minimum coordinates of the top, bottom, left and right centers of writing objects. In order to avoid missing those character pixels which are lying near or outside of the initial boundary, width and height of the boundary box are corrected by blob analysis. As one can see in figure 2 and 3 the cut out regions do not fit exactly to the words. The original area (see figure 3a) contains parts of the next write line and some unneeded dots. In figure 3b all blobs are computed and marked in red color. In figure 3c the small blobs are filtered out and the needless blobs outside a central band are colored in orange. Figure 3d shows the resulting area of the write line. Also in cases where high-quality commercial OCR programs could not locate text lines the above procedure gives clear referring to write lines. In some cases however also plant components in particular horizontal lying grasses become indicated as possible text regions. Such errors arise however also with commercial OCR programs. Thus Omnipage 17 recognized e.g. plant leaves and roots in an image as written text.

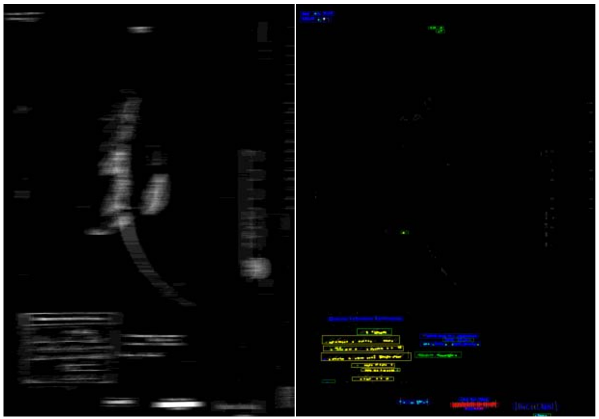


Fig. 2a,b: Writing objects merged to write lines

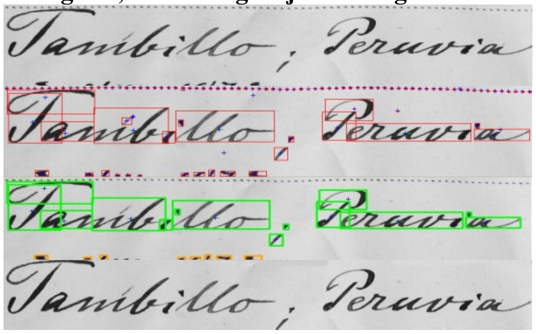

Tambillo, Peruvia
Tambillo, Peruvia
Tambillo, Peruvia
Tambillo, Peruvia

Figure 3a,b,c,d: Boundary boxes by blob analysis

3. Distinguishing writing styles

The image in figure 2b contains colored boxes with stamps, barcode, printed text, parts of plant and handwriting. In order to separate the different writings and to analyze the handwriting of different writers a statistical approach [10] was selected. The handwriting is seen as a texture with a steady structure of line elements all over the image. For the description of such a texture a suitable set of primitive elements has to be found whose frequency of occurrence is suited to distinguishing different writers to the greatest possible extent. The line segments of which the writing is composed can be taken as primitive elements of a handwriting specimen. Straight line segments may be obtained by the run lengths of pixel chains. The number and length of pixel chains is determined in eight different directions (see figure 5) and for each direction a frequency distribution is made. The features obtained by this shift-invariant transformation are nearly text independent as long as there is enough text at hand (about three to five handwriting lines). The feature vector furnishes information about the sloping position, size, regularity and roundness of the handwriting. The developed software can be imagined as a shredder (see figure 5) which is fed with 8 rotated documents (see figure 4). The feature vectors obtained by the described method have a very high dimension. As neighbored components of the vector are strongly correlated they are added to a certain degree so that only 8 features in each direction remain. Altogether a feature vector with 64 components results (see figure 6). By that the objects and the whole boundary box can be assigned to a color (see figure 2) where blue means printing, green means handwriting, red means barcode, turquoise means stamp and yellow means handwriting of Alexander von Humboldt. Experiments showed that the method is not only capable of distinguishing different writing styles but also different writers.

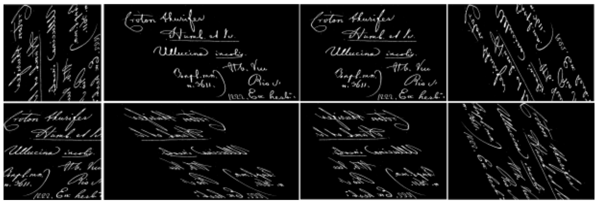


Figure 4: Original and transformed documents

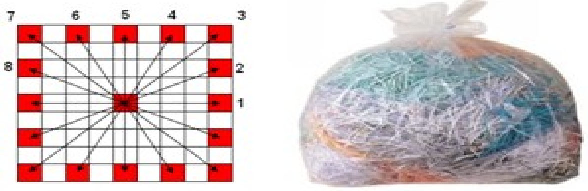


Figure 5: Eight directions for the shredder

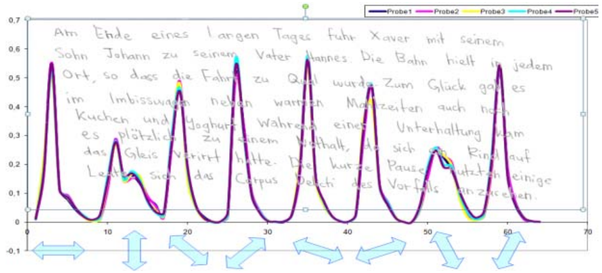


Figure 6: Feature vectors of 5 samples of 1 writer

3.1. Results with the IAM-Database

From the IAM-database [3] writers were chosen with at least 5 samples of each. 93 writers were found and so 93*5=465 sheets with handwriting were processed. With a simple nearest neighbor classifier and the leaving one out method the correct writer was found in 459 cases. Only 6 samples [25(5),32(3),37(2),40(1),52(4),88(2)]. were assigned to a wrong writer. This results in a recognition rate of 98,71%.

3.2. Results with historical material

From the Botanic Museum of Berlin 107 handwriting samples from 17 writers were chosen whereas 54 samples are from Alexander von Humboldt. Because the training set is of small size a nearest neighbor classifier was used. With the leaving one out method the correct writer was found in 102 cases. Only 5 samples were assigned to a wrong writer (see figure 7). This results in a recognition rate of 95,33% . In comparison with the IAM-database we deal with historical material which often contains overwritten words, crossed out and the paper is yellowed (see figure 8), molded (see figure 1) and has other artifacts..

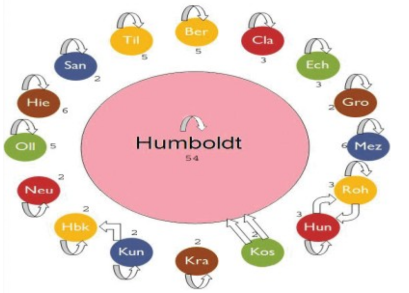


Figure 7: Results with historical material

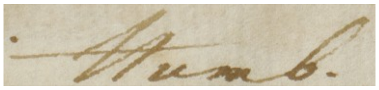


Figure 8: Sample of Humboldt's handwriting

4. Identification of writer

During on-line writing recognition e.g. with PDAs time series are used to read handwriting. In order to make a similar approach possible for already written handwriting on paper a half automatic software was developed which transfers handwriting into dynamic coordinate sequences. The handwritten character can be extracted out of a connected text. In order to recover the write line from documents whose writer is not present the following model serves: The writing is written as a groove in sand (see figure 9). A ball equipped with inertia rolls along the groove and reproduces the write line while keeping its last direction. If it arrives at a terminator point it will run back and try to deviate from the last way. In unclear situations the ball can be pulled manually with a „rubber band " (right mouse button) in the desired direction. If characters merge with the next text line the write line will be received as well.

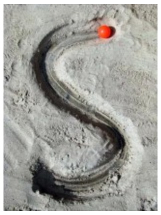


Figure 9: Ball with a groove in sand

The dynamic coordinate sequence is to a large extent independent of the used writing pen. In figure 10 a "W" from a specimen of handwriting and the coordinate sequence are compared.

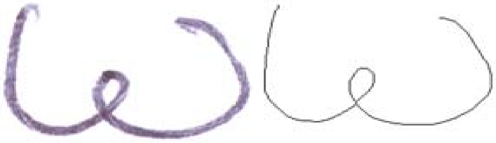


Figure 10: Comparison original and sequence of W

In order to compare writers by characters, the following procedures were explored:
-Approximation by Fourier series
-Approximation by Chebychev polynomials
-Approximation by Legendre polynomials

4.1. Fourier series

By a Fourier expansion a repetitive function can be represented as a set of sine and cosine functions, whose frequencies are integral multiples of the basic frequency ω=2π/T.

$$f(t)=\frac{a_0}{2}+\sum_{n=1}^{\infty}(a_n\cdot\cos(n\omega t)+b_n\cdot\sin(n\omega t))$$

The Fourier coefficients a_n und b_n can be computed by the Euler formulas. By the coefficients the character can be back-transformed. In figure 12 an original "W" and an approximation with different numbers of coefficients is shown.

4.2. Chebychev polynomials

Functions can be approximated by the use of Chebychev polynomials of the first kind with a very high accuracy. The polynomials are computed by the following formula:

$$T_n(x)=\cos(n\cdot ar\cos(x)),x\in[-1,1]$$

A polynomial $T_n(x)$ has exactly n zeros in the interval [-1,1].

In order to get the Chebychev coefficients needed in this application it is necessary to approximate the writing process of the characters by the following expression:

$$f(x)\approx\sum_{k=o}^{N-1}c_kT_k(x)-\frac{1}{2}c_0$$

N=number of coefficients

On the basis of the differences between the coefficients, the most similar characters can be determined.

4.3. Legendre polynomials

In the 3rd method for the determination of the character with the smallest deviation, the Legendre coefficients are used. In connection with the pertinent Legendre polynomials functions can be approximated. The ones generally represented look as follows:

$$P_n(x)=\frac{1}{\left(2^n n!\right)}\cdot\frac{d^n}{dx^n}\left[\left(x^2-1\right)^n\right]$$

A polynomial $P_n(x)$ has exactly n zeros in the interval [-1,1]. An approximation of the functions by the polynomials is done by

$$f(x)=\sum_{n=1}^{\infty}c_nP_n(x)$$

where c_n are the Legendre coefficients. They can be computed by

$$c_n=\frac{2n+1}{2}\int_{-1}^{1}f(x)P_n(x)dx$$

After the calculation a back transformed function can be produced by inserting the coefficients and the polynomials into the expression. An example of an approximation by the discrete Legendre transformation is given in figure 12 where "W" as original and as inverse transform is shown.

4.4. Splitting characters into time series

By the Fourier series the Chebychev and the Legendre polynomials functions can be approximated. As you can see in figure 11 the character “Z” is not a function. Up to three y-values for one x-value are existing. Therefore an approximation of the writing process cannot happen directly by one function. The problem is solved as the write line is split up into x and y-movement and afterwards each is separately discretely approximated.

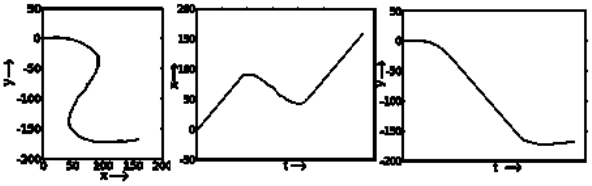


Figure 11: x-y-diagram, x-t-diagram, y-t-diagram

4.5. Reconstruction of characters

With the help of the characteristics of all used comparison methods the original writing process can be reconstructed again. The accuracy of these reproductions depends on the number of used features. Figure 12 shows the reproductions of a “W” by Legendre polynomials and Fourier series. With both methods the represented reproductions are numbered as follows:

Original character(1) , reconstruction by 64 features(2), 32 features(3), 16 features(4), 8 features(5), 4 features(6).

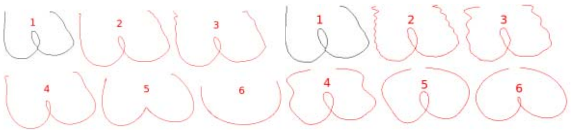
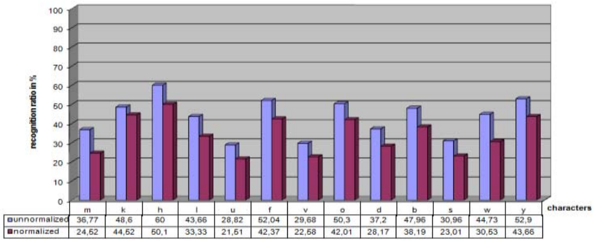


Figure 12: Reconstruction by Legendre polynomials and Fourier series

4.6. Results with the IAM-Database

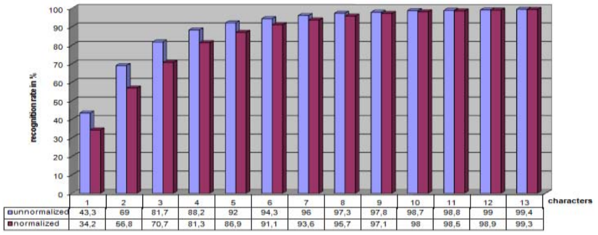
For the comparison with a large number of writers an international data base from Switzerland was chosen. The IAM-database [3] contains handwritings in English language with different texts. 6045 characters from 93 writers were extracted from the images (5 samples randomly chosen from each writer). The 13 characters m, k, h, l, u, f, v, o, d, b, s, w, y were selected (see figure 15). The characters possess different discriminatory abilities. In figure 13 the writer recognition rates using only one character is shown.



	m	k	h	l	u	f	v	o	d	b	s	w	y
▨unnormalized	36,77	48,6	60	43,66	28,82	52,04	29,68	50,3	37,2	47,96	30,96	44,73	52,9
■normalized	24,62	44,62	50,1	33,33	21,61	42,37	22,68	42,01	28,17	38,19	23,01	30,63	43,66

Figure 13: Recognition rates with one character

The combination of several characters promises a higher recognition rate by more information. For the comparison of the writers n characters of an unknown writer are selected. To each of the selected characters the most similar character of any of the well-known writers is assigned. The distances of the n most similar ones are added and a decision is made in favor of the writer with the minimum distance sum. It arises in figure 14 that the recognition rate with a higher number of used characters rises considerably. When combining the characters the procedures with the Legendre, Fourier and Chebychev coefficients prove themselves as almost equivalent. It is interesting that the size-normalized characters first supply worse recognition rates. With increasing character number they measure up with the not normalized characters.



	1	2	3	4	5	6	7	8	9	10	11	12	13
▨unnormalized	43,3	69	81,7	88,2	92	94,3	96	97,3	97,8	98,7	98,8	99	99,4
■normalized	34,2	56,8	70,7	81,3	86,9	91,1	93,6	95,7	97,1	98	98,5	98,9	99,3

Figure 14: ***Recognition rates with multiple characters***

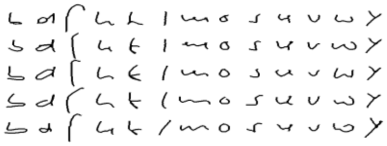


Figure 15: Thirteen characters of one writer (IAM)

4.7. Results with historical material

13 capital letters (261 samples altogether) were extracted from Alexander von Humboldt’s handwriting and compared with a writer set of 10 writers of a German database (IAM-database is in English and has not enough capital letters). The letters from Humboldt could be attached correctly to Humboldt in 96% of the cases. This result does not surprise however since the letters of Humboldt are very special and differ often clearly from the letters of other writers.

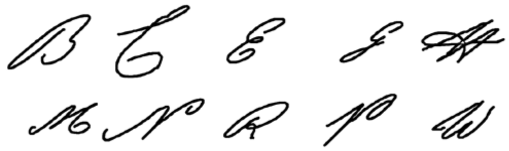
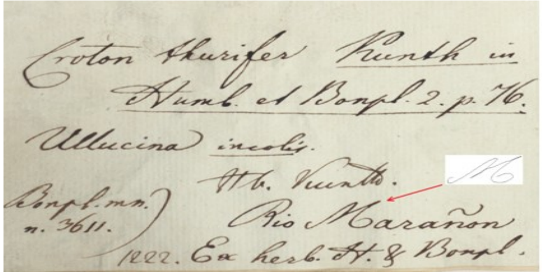


Figure 16: Capital letters of Humboldt



Croton thurifer Kunth in
Humb. et Bonpl. 2. p. 76.
Ullucina incola.
Kt. Kunth.
Bonpl. mss.
n. 3611.
Rio Marañon
1802. Ex herb. H. & Bonpl.

Figure 17: Humboldt's writing and an extracted M

5. Combining local and global method

From the IAM-database [3] writers were chosen with at least 5 samples of each. 93 writers were found and so 93*5=465 sheets with handwritings were processed. From each sheet 13 characters were extracted with the local method described in section 4. With a simple nearest neighbor classifier and the leaving one out method the correct writer was found in 459 cases. Only 6 samples [44(5),48(1),50(2),50(5),75(5),84(1)] were assigned to a wrong writer. The global method described in section 3 also mismatched only 6 but disjoint samples [25(5),32(3),37(2),40(1),52(4),88(2)]. Both methods are seen as two classifiers that have to be combined. Fortunately the both classifiers serve completely different features. Combination of classifiers can be performed on different system levels as described in [18]. A simple rank-based method was applied here. The strategy is based on the rank of result class in the j-best result list of each classifier. The ranks are added weighted by a cost function. As the recognition rates of both classifiers are similar both cost functions are set to 1. We come to a decision by the minimal sum of both ranks. The error rate falls to 2 of 465 samples (see figure 18).

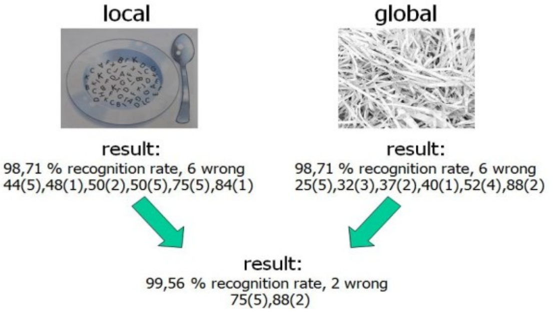


Figure 18: Combining local and global method

References

[1] Bensefia, A., Paquet, T., Heutte, L., A writer identification and verification system, Pattern Recognition Letters, vol. 26, issue 13, 2080-2092, 2005.

[2] Rath, T.M.,Manmatha, M., Word Image Matching Using Dynamic Time Warping, CVPR 2003, p 521.

[3] Marti, U., Bunke, H., A full english sentence database for off-line handwriting recognition, Proceedings of the 5. Int. Conference on Document Analysis and Recognition, Bangalore 1999, pp. 765-768.

[4] Marti, U., Messerli, R., Bunke, H., Writer Identification Using Text Line Based Features, Proc. Of the 6th International Conference on Document Analysis and Recognition, Seattle, USA, 2001, pp. 101-105.

[5] Niels, R., Grootjen, F., Vuurpijl, L., writer identification through information retrieval: the allograph weight vector, Proceedings of the 11. Int. Conference on Frontiers in Handwriting Recognition, Montreal, 2008.

[6] Srihari, S. , Arora, H., Lee, S., Individuality of handwriting, J. of Forensic Sciences, 47(4):1.17, July 2002

[7] Schlapbach, Andreas; Bunke, Horst: Off-line Handwriting Identification Using HMM Based Recognizers, 2004, publications Uni Bern

[8] Schomaker, L., Bulacu, M., Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script, in IEEE Transactions of Pattern Analysis and Machine Intelligence, vol. 26, no. 6, pp. 787-798, 2004.

[9] Rousseau L.,Anquetil E., Camillerapp J.. What knowledge about handwritten letters can be used to recover their drawing order?.Proceedings of the 10th International Workshop on Frontiers in Handwriting recognition (IWFHR 2006), Octobre 2006.

[10] Steinke, K.-H. , Gehrke, M., Dzido, R., Writer Recognition by Combining Local and Global Methods, International Congress on Image and Signal Processing, Tianjin China, October 2009

[11] Siddiqi, I., Vincent, N., Combining global and local features for writer identification, Proceedings of the 11. Int. Conference on Frontiers in Handwriting Recognition, Montreal, 2008

[12] Steinherz,T., Doermann,D., Rivlin,E., Intrator,N., Offline Loop Investigation for Handwriting Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 193-209, February, 2009

[13] Sobottka, K., Bunke, H., Kronenberg, H., Identification of text on colored book and journal covers, Proceedings of the 5. Int. Conference on Document Analysis and Recognition 1997

[14] Kanungo, T., What fraction of images on the web contain text, Proceedings of Web Document Analysis, 2001

[15] Wu, V., Manmatha, R., Riseman, E. M., Finding text in images, Proc. ACM Int. Conf. Digital Libraries 1997

[16] Lienhart, R., Stuber, F., Automatic text recognition in digital videos, Proceedings of the SPIE Image and Video Processing IV 1996

[17] Chen, X., Yuille, A., Detecting and Reading Text in Natural Scenes, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004

[18] Kuncheva, L. I., Combining Pattern Classifiers. Methods and Algorithms, Wiley, 2004