Writer Retrieval and Writer Identification using Local Features

Stefan Fiel and Robert Sablatnig
Computer Vision Lab
Institute of Computer Aided Automation
Vienna University of Technology
fiel@caa.tuwien.ac.at

***Abstract*—Writer identification determines the writer of one document among a number of known writers where at least one sample is known. Writer retrieval searches all documents of one particular writer by creating a ranking of the similarity of the handwriting in a dataset. This paper presents a method for writer retrieval and writer identification using local features and therefore the proposed method is not dependent on a binarization step. First the local features of the image are calculated and with the help of a predefined codebook an occurrence histogram can be created. This histogram is compared to determine the identity of the writer or the similarity of other handwritten documents.**

The proposed method has been evaluated on two datasets, namely the IAM dataset which contains 650 writers and the TrigraphSlant dataset which contains 47 writers. Experiments have shown that it can keep up with previous writer identification approaches. Regarding writer retrieval it outperforms previous methods.

***Keywords*-writer retrieval, writer identification, local features**

I. Introduction

The objective of writer identification is to determine the writer of a handwritten text among a number of known writers. A database of specific features for each writer has to be built up in advance and when identifying a new text features are calculated and compared to the ones stored in the database. The writer of the document in the database with the highest similarity is then assigned as writer for the new text. In contrast to this writer retrieval addresses the problem to obtain all documents of one writer out of a set of documents. Therefore a ranking of pages according to the similarity of the handwritings to the writing of a query page is generated. This task can be used for example to retrieve all documents of one writer out of an archive.

Currently the methods for writer identification can be divided into two approaches: the first approach analyzes the characters themselves and the second approach uses textural features of the handwriting. In forensics the writer identification is done by analyzing the style of the characters. For this analysis it is necessary that the foreground has to be seperated from the background in the images, which makes the results of the writer identification dependent of the binarization algorithm. Additional these algorithms have problems with faded out or blurred ink. When using textural features for writer identification no separation of foreground is necessary, thus making it independent of a binarization step. The drawback is that more text is needed for the identification.

Brink et al. [1] examined writer identification algorithms to show how much handwritten text is needed for an identification. They showed that when using string features 100 characters are sufficient, when using less powerful features a minimum of 200 characters are required.

This paper presents an approach based on textural features for writer retrieval, which can also be used for writer identification. First local features are calculated on the input image. Afterwards a histogram is generated using the bag of words approach. This histogram can then be used to either identify the writer or get the documents of one particular writer.

This paper is organized as follows: Section II gives a brief description of the current state of the art. In Section III the methodology of this approach is described. Section IV presents the experiments and the results. Finally, a short conclusion is given in Section V.

II. Related Work

Writer identification methods can be divided into two main approaches. On the one hand methods that use features which are based on the characters and on the other hand algorithms which use textural features. If the features are calculated on the characters the image needs to be segmented first. Marti et al. [2] are using features extracted from the handwritten lines of text. These features comprise width, slant, and the three heights of the writing zones (descender height, ascender height, and the height of the writing itself). Using a neural network as classifier a recognition rate of 90.9 % is achieved. Hertel et al. [3] introduced new features like connected components, enclosed region and the lower and upper contour of the writing. In an experiment a recognition rate of 99.6% is shown.

Bulacu et al. [4] used the contour-hinge, the writer-specific grapheme emission and the run-length for writer identification. They achieved a result of 89 % using $k-NN$ for classification.

Schlapbach et al. [5] applied a Hidden Markov Model (HMM) based recognizer for the writer identification. For each writer a HMM is trained and the system returns the

identity of the text with the highest ranked score. When using the 6 nearest neighbors an identification rate of 97 % is reached.

Tan et al. [6] proposed a method to estimate statistical distributions of character prototypes on an alphabet basis. These distributions model the unique handwriting styles of the writers. With a Fuzzy C-Means approach for classification an identification rate of 96.7% is achieved.

Hiremath et al. [7] presented a binarization free approach. The writing is assumed as texture image and thus the writer identification is a texture classification. In the subband images of the wavelet transform co-occurrence matrices are computed. This is done for 8 directions. When dealing with 30 writers at a time the classification accuracy is 88%. Du et al. [8] proposed a method using wavelet domain local binary pattern features for writer identification of chinese handwriting. On a database with 50 writers an identification rate of 90% is achieved when using a hitlist of size 4.

For writer retrieval Atanasiu et al. [10] proposed a method using 10 perceptual features from script orientation. They evaluated the efficiency of each feature for the task of writer retrieval. When using several features and choosing the best one for each query document all documents of a writer are retrieved by selecting 70% of the database documents.

The Document Image Binarization Contest [11] showed that the binarization of documents is still a challenging task. An incorrect binarization leads to wrong features at character level, thus a method without this step is proposed.

III. Methodology

The proposed method also assumes that the writing is a texture image, thus a binarization step is not necessary. The benefit is that the writer identification or writer retrieval does not depend on a binarization algorithm. The two methods are presented in Figure 1. The task of writer identification is illustrated in Figure 1 a), whereas the challenge of writer retrieval is showed in Figure 1 b). For the writer retrieval the features of all documents in the dataset have to be generated and the query document has to be compared with every document in the dataset and the χ^2 distance between the two histograms is used as similarity measure. The output is a ranking of the similarity of the documents in the dataset. For the writer identification a database of documents where the writer is known has to be created. To classify a new document it is compared to all documents stored in the database and the writer of the document with the smallest distance is assigned as writer for the new document. Both methods have in common that a codebook based on bag of words [12] has to be generated. This is done by calculating the Scale Invariant Feature Transform (SIFT) features [13] on a various pages of handwriting. These features are then clustered using k-means and the cluster centers form the codebook. The different steps for both methods are now described in more detail.

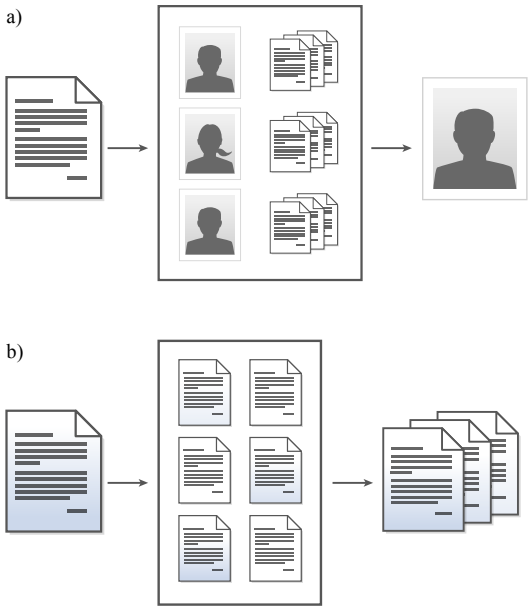

Figure 1. a) illustrates the task of writer identification. A new document is compared with all documents in the database and the result is the identity of the writer with the smallest difference. b) shows the task of writer retrieval. A new document is compared with all documents and a ranking of the similarity of the handwriting is generated.

A. Writer Retrieval

For the writer retrieval the SIFT features have to be calculated on the normalized images. The features for one image are then compared with the cluster centers. The most similar cluster center according to the euclidean distance is searched. With the occurrences of the cluster centers a histogram for each image is built up. These steps can either be calculated in advance or just in time. For the experiments 300 cluster centers are used, which are determined empirically. Figure 2 shows the generation of the histogram on two sample images of two different writers.

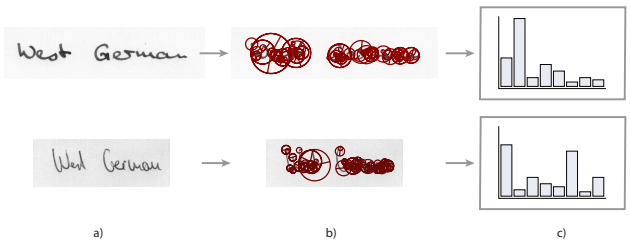

Figure 2. a) two words of a sample page by two different writers b) the calculated SIFT features c) a sample histogram with 8 bins of occurrences of the cluster centers.

When searching the documents with the most similar handwriting for a new image the histogram as described above is calculated for this image. Afterwards it is compared with each histogram of the documents in the dataset. This

comparison is done using the χ^2- distance since experiments have shown that it lead to the best results in comparison with the euclidean and earth movers distance. The distances can then be sorted and a ranking for the similarity is created.

B. Writer Identification

For a writer identification task the writers have to be known in advance. For each writer at least one document is taken to form a database. For all documents in the database the SIFT features are calculated and the histogram of occurrences of the cluster centers in the codebook is created. For the experiments also 300 cluster centers are used, which were also determined empirically.

When searching for the identity of a writer again the histogram is built up and compared to the ones in the database. For this comparison also the χ^2-distance is used since it performed the best. Using a nearest neighbor classification the identity of the writer can be determined.

IV. Experiments and Results

For the experiments two datasets were used. The first dataset is the IAM dataset by Marti and Bunke [14]. It contains the handwriting of 657 different people and each person has written up to 59 pages (the average is 2.3 pages, 356 writers have only one document) in English. In total the dataset has 1539 images. Figure 3 shows one sample image of the IAM datset.

The second dataset is the TrigraphSlant dataset by Brink et al. [15] which consists of 188 scanned images of handwritten pages written by 47 writers. The writers have written four pages in Dutch each: two with the natural handwriting and the other two with the maximal slant of the handwriting to the left respectively to the right, so for our experiments only the two pages with the natural handwriting are taken into account. Figure 4 shows one sample image of the TrigraphSlant dataset.

For the writer retrieval and writer identification a codebook has to be generated. In our experiments the codebook is always created with all features of the dataset which is not used to ensure independence between the codebook and the test dataset. When the codebook is created with the TrigraphSlant dataset also the pages with the unnatural slant are taken into account.

A. Experiments and Results for Writer Retrieval

First experiments for the writer retrieval have been carried out. For each document a ranking of the most similar documents is created. To evaluate a query document the rankings of all other documents in the dataset are generated. Afterwards it is checked whether the first N documents of the ranking are written by the same person as the query document. The number of documents which are checked depends on the number of documents in the dataset from the particular writer of the query document. If e.g. 10 documents

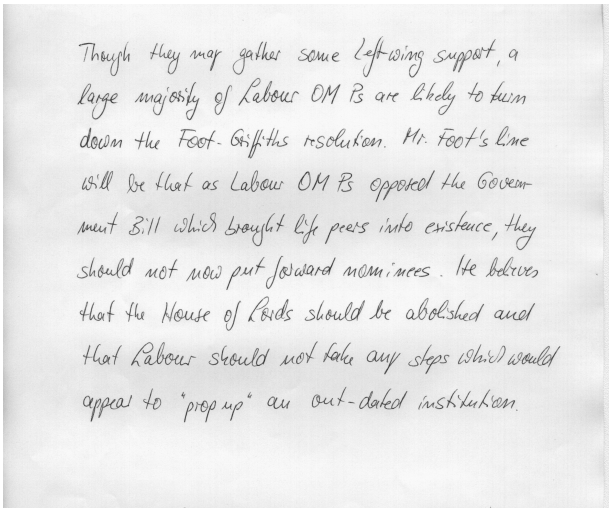

Figure 3. One sample image of the IAM dataset.

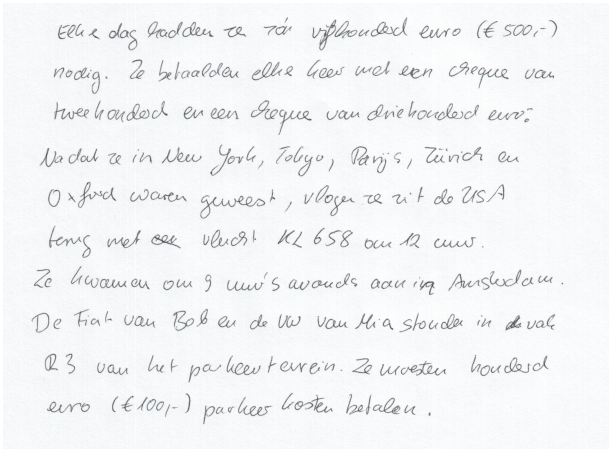

Figure 4. One sample image of the TrigraphSlant dataset.

of a specified writer exist in the dataset, the writer of the first 9 documents in the ranking are compared to the one of the query document. If a writer has only one document in the dataset, this document is skipped as query document but the document remains in the dataset for the other tests.

This means that for the TrigraphSlant database where every writer has 2 documents in natural slant for each document only the first document in the ranking is considered, so for the complete dataset 94 documents in the rankings are checked if they are correct. Since in the IAM dataset the number of pages per writer is not equally distributed, in total 8102 documents in the rankings are checked.

The first experiment is on the IAM dataset. Each document is used as query document and as described above the number of correct ranked documents is evaluated. The result of this experiment is 93.1%, this means that 7543 documents are correctly in the first neighbors, leaving 559 documents in the ranking which are more similar to the query document

then other documents of the particular writer. Figure 5 shows the ranking of 99 sample images of the IAM dataset. The query document is written by Writer 1 and it can be seen that the distance of the other documents of writer 1 has a smaller distance than to the other writers and that the documents of the other writers form clusters.

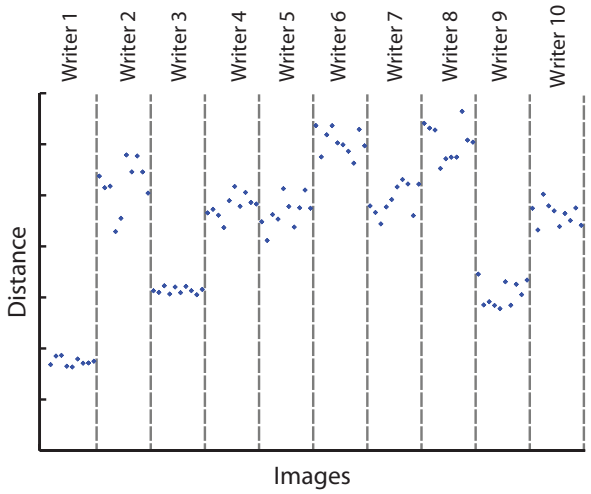

Figure 5. Ranking of 99 sample images when compared to one document of writer 1.

The second experiment is carried out using the Trigraph-Slant database. For each document the ranking is created and it is evaluated if the other document of the writer is ranked at the first place. The result of this experiment is 98.9% which means that for one document the second document of the same writer has not been found.

The last experiment for writer retrieval is a comparison to the results of Atanasiu et al. [10] who are also using the IAM dataset. For every writer, which has more than one document in the dataset, a query document is chosen. The query documents are compared to all other documents in the dataset and with a nearest neighbor classification the other documents of the writers are found. Figure 6 shows the comparison of the two methods. The ordinate shows the percent of the retrieved documents for all the writers of the query documents. The abscissa shows the top-N which were taken into account. The dotted red line is the "upper limit" of Atanasiu et al. [10] which means that for each query document the best feature is chosen whereas the solid blue line is their best overall feature. The solid red line shows the performance of the proposed method. Since one writer has 59 documents in the dataset, 100% can be achieved not until 58 neighbors are regarded (vertical dashed line). At this point the proposed methods has a retrieval rate of 97.2%.

B. Experiments and Results for Writer Identification

For the experiments for the writer identification a database of known writers have to be determined. In our experiment

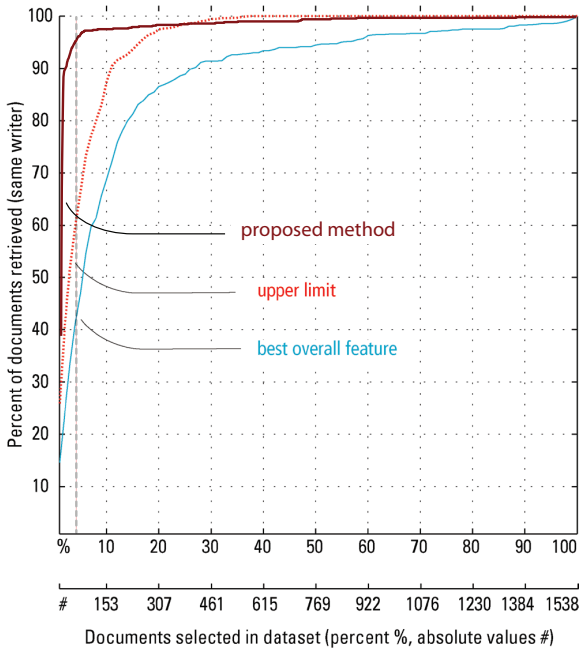

Figure 6. Comparison of the evaluation of the proposed method (dark red line) and Atanasui et al. [10]. The dotted rot red line is the upper limit of Atanasui et al. and the blue line is the best overall feature of Atanasui et al. The y axis is the retrieval rate and the x axis shows the number of neighbors which are taken into account. 100% retrieval rate cannot be achieved until 58 neighbors are taken into account (vertical dashed line).

the first document of each writer in the dataset is taken. For the classification a nearest neighbor classifier is used. Figure 7 shows the identification rate regarding a different amount of neighbors. For 300 cluster centers when only the closest neighbor is taken the identification rate is 90.8%, when regarding the top-5 the rate raises to 96.7% using a soft criterion. When 10 neighbors are taken into account, the identification rate is 97.5%.

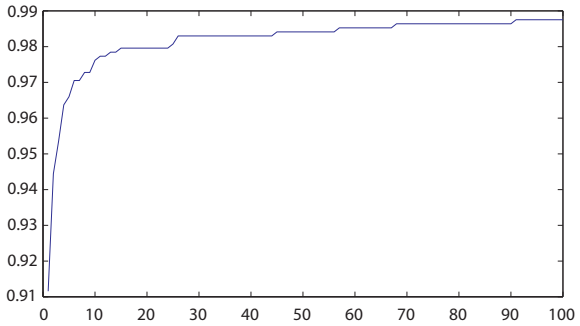

Figure 7. Writer identification rate of the IAM dataset. The y axis shows the identification rate and the x axis shows the top-N used.

On the TrigraphSlant database also the first document of the writer are taken for the database which leaves only one query document for each writer. Using a nearest neighbor

classifier with regarding only one neighbor the identification rate is 98.9%, which means that one document has not been assigned to the correct writer. It is the same document as in the experiment for writer retrieval.

V. Conclusion

A method for writer retrieval and writer identification has been presented in this paper. The difference between writer retrieval and writer identification is, that for the writer identification a database of documents where the writer is known has to be created and the algorithm assigns a writer to an input document. For writer retrieval a ranking according to the similarity of the handwriting of documents is created. The approach presented uses SIFT features and bag of words. First a codebook has to be generated and according to this codebook a occurrence histogram of the cluster centers of each image can be generated. These histograms are than compared. For the writer retrieval the distance of the histogram of the new image and all histograms in the dataset are calculated using the χ^2-distance and the ranking is the corresponding distance.

For the writer identification first a dataset of documents of known writers has to be created. The histogram of occurrences of a new document is then compared to the ones in the dataset, again, using the χ^2-distance. With a nearest neighbor classifier the identification of the writer can be determined.

The advantage of this method is that the characters do not have to be binarized. A bad binarization, which can occur due to the faded out ink or due to low contrast of old documents, will lead to wrong features when they are calculated on character level.

Additional, since both the IAM dataset is in English and the TrigraphSlant dataset is in Dutch it has been shown that the proposed method is language invariant. The experiments showed that the proposed method can keep up with previous approaches for writer identification and outperforms previous work for writer retrieval.

Acknowledgment

The authors would like to thank the Fraunhofer-Institute for Production Systems and Design Technology (IPK), Berlin for supporting the work.

References

[1] A. Brink, M. Bulacu, and L. Schomaker, "How much handwritten text is needed for text-independent writer verification and identification," in *19th International Conference on Pattern Recognition*, 2008, pp. 1 –4.

[2] U.-V. Marti, R. Messerli, and H. Bunke, "Writer identification using text line based features," in *Proceedings. Sixth International Conference on Document Analysis and Recognition*, 2001, pp. 101 –105.

[3] C. Hertel and H. Bunke, "A set of novel features for writer identification," in *Audio- and Video-Based Biometric Person Authentication*, ser. Lecture Notes in Computer Science, J. Kittler and M. Nixon, Eds. Springer Berlin / Heidelberg, 2003, vol. 2688, pp. 1058–1058.

[4] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 701 –717, 2007.

[5] A. Schlapbach and H. Bunke, "A writer identification and verification system using HMM based recognizers," *Pattern Analysis and Applications*, vol. 10, pp. 33–43, 2007.

[6] G. X. Tan, C. Viard-Gaudin, and A. C. Kot, "Automatic writer identification framework for online handwritten documents using character prototypes," *Pattern Recognition*, vol. 42, no. 12, pp. 3313 – 3323, 2009.

[7] P. Hiremath, S. Shivashankar, J. Pujari, and R. Kartik, "Writer identification in a handwritten document image using texture features," in *International Conference on Signal and Image Processing (ICSIP)*, dec. 2010, pp. 139 –142.

[8] L. Du, X. You, H. Xu, Z. Gao, and Y. Tang, "Wavelet domain local binary pattern features for writer identification," in *20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 3691 –3694.

[9] A. Gordo, A. Fornés, E. Valveny, and J. Lladós, "A bag of notes approach to writer identification in old handwritten musical scores," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, ser. DAS '10. New York, NY, USA: ACM, 2010, pp. 247–254. [Online]. Available: http://doi.acm.org/10.1145/1815330.1815362

[10] V. Atanasiu, L. Likforman-Sulem, and N. Vincent, "Writer retrieval - exploration of a novel biometric scenario using perceptual features derived from script orientation," in *International Conference on Document Analysis and Recognition*, 2011, pp. 628 – 632.

[11] I. Pratikakis, K. Ntirogiannis, and B. Gatos, "ICDAR 2011 Document Image Binarization Contest (DIBCO 2011)," in *11th International Conference on Document Analysis and Recognition, 2011. ICDAR '11.*, jul. 2011, pp. 1506 –1510.

[12] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," *In Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22, 2004.

[13] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] 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, pp. 39–46, 2002.

[15] A. Brink, R. Niels, R. van Batenburg, C. van den Heuvel, and L. Schomaker, "Towards robust writer verification by correcting unnatural slant," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 449 – 457, 2011.