ICFHR 2014 Competition on Handwritten Digit String Recognition in Challenging Datasets (HDSRC 2014)

Markus Diem, Stefan Fiel, Florian Kleber and
Robert Sablatnig
Computer Vision Lab
Vienna Univ. of Technology
1040 Vienna, Austria
{diem,fiel,kleber,sab}@caa.tuwien.ac.at

Jose M. Saavedra[1], David Contreras[1] and
Juan Manuel Barrios [1], Luiz S. Oliveira[2]
[1]*Orand S.A. (Chile) and*
[2]*Federal University of Parana (UFPR, Brazil)*
jose.saavedra@orand.cl

***Abstract*—This paper presents the results of the HDSRC 2014 competition on handwritten digit string recognition in challenging datasets organized in conjunction with ICFHR 2014. The general objective of this competition is to identify, evaluate and compare recent developments in Western Arabic digit string recognition with varying length. In addition, this competition introduces two new challenging datasets for benchmarking. We describe competition details including the datasets and evaluation measures used, and give a comparative performance analysis of six (6) participating methods along with a short description of the respective methodologies.**

***Keywords*-digit; string; recognition; competition;**

I. Introduction

Due to the high variability of handwriting, recognition of unconstrained handwriting is still considered an open research topic in the document analysis community. The ICDAR 2013 competition on handwritten digit recognition (HDRC 2013) [1] has shown that single handwritten Arabic digits can be recognized with a precision of 97.74% (99.33% including second guess). However, recognizing handwritten digit strings introduces new challenges since we have to deal with connected digits. Moreover, in real environments we will have to face the innate noise of the images (document layout, or noisy strokes). Hence, a new framework for benchmarking, i.e. two new freely available real world datasets along with objective evaluation measures are provided in order to assess the performance of Western Arabic digit string recognition approaches.

We also present two challenging datasets of handwritten digit strings which are freely available[1]. The first one was collected amongst students of the Vienna University of Technology showing high variability about length and handwriting style. The second one was obtained from the Courtesy Amount Recognition field (CAR) of real bank checks containing innate noise and variability.

This paper is organized as follows: a detailed description of the two proposed datasets is presented in Section II. The description of the submitted methods is presented in Section III and the evaluation metrics together with the evaluation analysis are shown in Section IV. Finally, conclusions are drawn in Section V.

II. Digit String Databases

The first database, named Computer Vision Lab (CVL) Handwritten Digit String (HDS), has been collected mostly amongst students of the Vienna University of Technology and consists of about 300 writers, female and male alike. The variability of writers allows us to get high variability with respect to handwriting styles. The CVL HDS database is composed of 7960 images, from which 1262 images have been proposed for training and the other 6698 images, for testing. The publicly available dataset containing segmented digits (HDRC 2013) [1] were extracted from this database. An example of a CVL HDS image is shown in Figure 1.

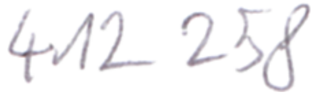

Figure 1. Example digit string of the CVL HDS database.

The second database, named ORAND-CAR, consists of 11719 images obtained from the CAR field of real bank checks. The CAR images are composed of a combination of special numeric symbols representing a monetary value. The ORAND-CAR images come from two different sources which give the images different characteristics. These characteristics are mainly related to the image quality, kind of noise and handwriting style. Therefore, considering the two different sources, the ORAND-CAR database is split into two subsets: ORAND-CAR-A, that comes from an Uruguayan bank and ORAND-CAR-B that comes from a Chilean bank.

The ORAND-CAR-A database consist of 2009 images for training and 3784 images for testing and the ORAND-CAR-B database consists of 3000 images for training and

[1] http://www.orand.cl/en/icfhr2014-hdsr/

2936 images for testing. Figures 2 and 3 show examples of gray-scale CAR images of the ORAND-CAR database.

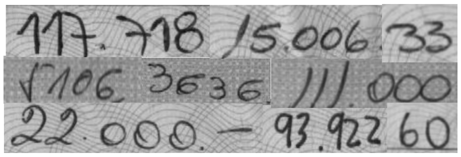

Figure 2. A sample of images form ORAND-CAR-A database.

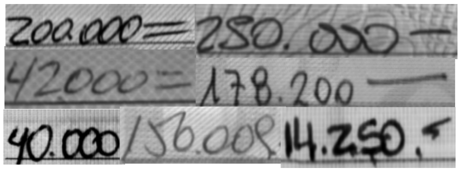

Figure 3. A sample of images form ORAND-CAR-B database.

As it can be noticed, real CAR images bring with a variety of noise like background patterns, check layout, or user strokes. In addition, the natural variability of the handwritten text represents another challenge that must dealt with.

The CVL HDS database has, in contrast to the ORAND-CAR database, *no* background noise. Thus, the combination of both datasets leads to a database which comprises the described different scenarios. Figure 4 shows the results (precision 1st guess) of all participants on both ORAND-CAR subsets and on the CVL HDS database. White vertical strokes illustrate the performance of a method and black vertical strokes represent the median precision on a dataset. It can be noted that ORAND-CAR-B is the most challenging database with a median precision of 0.44. In addition, the participants achieve the highest median precision (0.59) on the CVL HDS database. For ORAND-CAR-A, the median precision was 0.51.

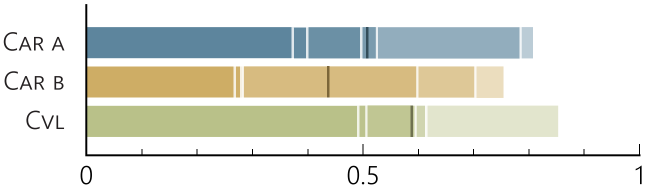

Figure 4. Overall performance, in terms of median precision, achieved on the different databases. Each transparent bar indicates the performance of a participating method whose end is marked by a white line. Dark lines denote the median performance of all participating methods on each database.

Figure 5 shows the distribution of string lengths in the different databases. This property is important to track differences between the performance metrics which are introduced in Section IV. It is illustrated, that the CVL database has on average the longest strings (6.1 digits) while the ORAND-CAR-A contains most short strings (4.5 digits).

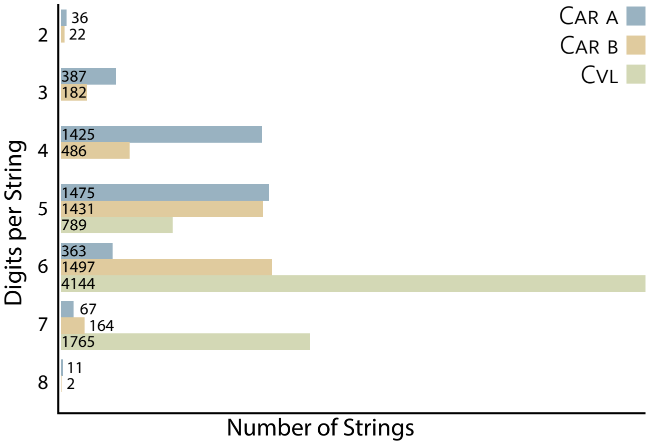

Figure 5. Distribution of string lengths with respect to the three different databases.

III. Participants and Methods

Five (5) research groups have participated with six (6) methods for the recognition of handwritten digit string. One group submitted two different algorithms.

1. Beijing: National Laboratory of Pattern Recognition (NLPR) Institute of Automation of Chinese Academy of Sciences, Beijing, China (Yi-Chao Wu, Fei Yin, Chang Zhong, Cheng-Lin Liu)

The input image is firstly classified as a simple or a complex image according to the distribution of pixel values using a neural network. For a complex image, we use two MLP neural networks for cleaning and enhancing the image, and then use the Sauvola method to binarize it. A simple image is binarized using the Otsu's method. In the binarized image, connected components are firstly slant-corrected, then those heavily overlap in horizontal direction are merged, finally the components with large width (relative to the estimated string height) or abnormally large width/height ratio are regarded as potential touching patterns and are split by upper/lower profile curve analysis to generate vertical cuts [2]. After that, the string image is represented as a sequence of primitive image segments. Consecutive primitive segments are combined to generate candidate character patterns, forming a segmentation candidate lattice. Each path from the start node to the end node corresponds to a candidate segmentation. The candidate patterns are classified using a polynomial classifier with gradient direction histogram feature. Via confidence transformation, the character classifier and binary class-dependent geometric model [3] are combined together to evaluate the segmentation path. A beam search algorithm [4] is used to find the optimal path with minimum cost or maximum score.

2. Pernambuco: Universidade de Pernambuco, Santo Amaro, Brasil (Byron Leite and Cleber Zanchettin)

This method was built over the last decade taking into account real data extracted from Brazilian bank checks. The architecture of this system was designed to allow the combination of multiple classifiers. In order to cope with some sort of noise, an appropriate pre-processing algorithm that detects non-digit structures [5] was developed. In case of gray level images, the system chooses the best thresholding level for a set of pixels using an MLP neural network [6]. Recently, two hybrid classifiers were aggregated in the system in order to improve the accuracy rate. One classifier is based on a hybrid model combining a k-NN model with an SVM approach which improves the classification accuracy in high similarity cases, such as digits 5 and 6. The last classifier integrated is based on the combination of MDRNN [7] and SVM methods [8]. With this last classifier, the authors improve the accuracy of the system recognizing highly connected digits without a segmentation step.

3. Shanghai: ECNU-SRI Joint Lab for Pattern Analysis and Intelligent System, Shanghai Research Institute of China Post, Shanghai, China (Shujing Lu and Yue Lu)

In this proposal, the digit characters are segmented based on adaptive binarization and connected component analysis firstly. After the local directional features are extracted, character classifiers, including SVM, SOM and BP network, are used to recognize the segmented digit images and their recognition results for each digit image are ordered on confidences. Finally, the TOP-3 guesses for the digit string image are built up in terms of the ordered recognition results.

4. Singapore: Institute for Infocomm Research, Singapore (Su Bolan and Lu Shijang) and School of Computing, National University of Singapore (Zhang Xi and Tan Chew Lim)

In this approach, the input image is first pre-processed for feature extraction. Then, the pre-processed image is converted into sequential feature vectors using the HOG approach. Finally, the Recurrent Neural Network model is applied to obtain the recognition result without segmentation.

5. Tébessa I: LAMIS Laboratory University of Tébessa, Algeria, (Abdeljalil Gattal) and Speech Communication and Signal Processing Laboratory, Faculty of Electronics and Computer Science, University of Sciences and Technology Houari Boumedienne, Bab-Ezzouar, Algiers, Algeria (Youcef Chibani)

This method uses a forward segmentation strategy based on oriented sliding window [9], [10]. It uses two complementary sets of structural features (contour and skeletal points) and the concept of oriented sliding window. Radon transform is applied to handwritten connected digits in order to get orientation angle of the sliding window oriented. The histogram of vertical projection is applied to the digit image to detect digits that are not overlapped or not connected by the recognition stage. When the digit image is not recognized, the image contains more than one digit. In this case, the algorithm based on oriented sliding window is applied. Its workflow is to:

- Generate segmentation feature points (BPs and IPs).
- Set sliding window on the IPs positions.
- Crossing oriented window around IPs with sloping angle for finding correctly cutting path.
- Scan all the possible relationships of BPs (Bases points) and IPs (Interconnection points).
- Evaluate the correct segmentation path using a SVM.

The segmentation verification using the global decision module allows the rejection or acceptance of the processed image. For each image the three best responses (Top1, Top2 and Top3) are presented:

- Top1: Recognition with an original image;
- Top2: The same steps of Top1, but in this case, the sliding window in the reverse direction (right to left) is used.
- Top3: The same steps of Top1, but a normalized image is used.

6. Tébessa II: LAMIS Laboratory University of Tébessa, Algeria, (Abdeljalil Gattal and Chawki Djeddi) and Speech Communication and Signal Processing Laboratory, Faculty of Electronics and Computer Science, University of Sciences and Technology Houari Boumedienne, Bab-Ezzouar, Algiers, Algeria (Youcef Chibani)

The description of this method is similar to the previous (*Tébessa I*), but various orientation angles of oriented sliding window [9] selected in the range $\theta = \{-\alpha\,^\circ, 0\,^\circ, +\alpha\,^\circ\}$ are used. The results obtained for various angles $\alpha = \{-6\,^\circ, 0\,^\circ, 6\,^\circ\}$. Responses for each image amount into three responses (Top1, Top2 and Top3) are presented:

- Top1: The result is obtained when the orientation angle is fixed to $\alpha = 0°$.
- Top1: The result is obtained when the orientation angle is fixed to $\alpha = -6°$.
- Top3: The result is obtained when the orientation angle is fixed to $\alpha = 6°$.

IV. Evaluation

We use two evaluation metrics, a hard metric and a soft metric. In case of the hard metric, we measure the **precision** (recognition rate). That is, the number of correctly recognized digit strings divided by the total number of strings. The best three guesses are evaluated (TOP-3).

In addition, considering that the test images may be drastically affected by noise, the recognition rate might be very low for certain methods. To deal with this problem,

Submission	Guesses	CAR A	CAR B	CVL	Mean
Tébessa I	TOP-1	0.3705	0.2662	0.5930	0.4099
	TOP-2	0.4559	0.3401	0.6575	0.4845
	TOP-3	0.4720	0.3568	0.6690	0.4993
Tébessa II	TOP-1	0.3972	0.2772	0.6123	0.4289
	TOP-2	0.4477	0.3137	0.6527	0.4714
	TOP-3	0.4818	0.3411	0.6824	0.5018
Singapore	TOP-1	0.5230	0.5960	0.5040	0.5410
	TOP-2	0.6180	0.6770	0.6060	0.6337
	TOP-3	0.6540	0.7130	0.6540	0.6737
Pernambuco	TOP-1	0.7830	0.7543	0.5860	0.7078
	TOP-2	0.8916	0.8746	0.6850	0.8171
	TOP-3	0.9199	0.9009	0.7234	0.8481
Beijing	TOP-1	0.8073	0.7013	0.8529	0.7872
	TOP-2	0.8634	0.7638	0.9128	0.8467
	TOP-3	0.8697	0.7779	0.9189	0.8555
Shanghai	TOP-1	0.4950	0.2809	0.4893	0.4217
	TOP-2	0.5378	0.3120	0.5400	0.4633
	TOP-3	0.4950	0.2809	0.4893	0.4849

Table I
PRECISION ACHIEVED FOR EACH SUBMITTED METHOD CONSIDERING UP TO TOP-3 GUESSES.

we also propose a soft metric that evaluates how close a resulting answer is to a target amount. The proposed soft metric is based on the **Levenshtein distance** (LD), which is also known as the *edit distance*. In particular, we use the **Normalized Levenshtein Distance** (NLD) to avoid any bias with respect to the string length.

Let a_T be a target string (GT) and a_R be the corresponding recognized string, the NLD metric is computed by:

$$\mathrm{NLD}(a_T, a_R) = \frac{\mathrm{LD}(a_T, a_R)}{|a_T|}, \tag{1}$$

where $|a_T|$ is the length of the string a_T and LD is the Levenshtein distance. In contrast to the hard metric only the best response (TOP-1) is evaluated.

We also define the **Average Normalized Levenshtein Distance** (ANLD) as below:

$$\mathrm{ANLD} = \frac{\sum_{i=1}^{T} \mathrm{NLD}(a_T^i, a_R^i)}{T}, \tag{2}$$

where T is the number of test digit strings. The ANLD will provide the final score of a method submitted. The ANLD is an inverse performance metric, i.e. a low value indicates high performance, while a high value indicates low performance. Software crashes are labeled as false recognition.

Considering the evaluation metrics discussed above, the precision achieved for each participating method is presented up to the third guess in Table I. The last column is the mean precision, with respect to the three databases.

Figure 6 shows the performance of the six methods on the ORAND-CAR-A database. Analyzing the first guess, two methods, *Pernambuco* and *Beijing*, clearly outperforms the rest. *Beijing* achieves a precision of 0.807 while *Pernambuco* achieves 0.783.

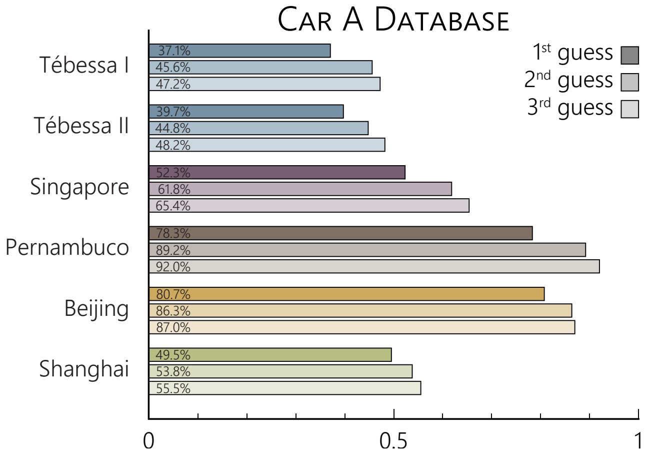

Figure 6. Precision of all participants on the CAR A database.

In the case of ORAND-CAR-B, Figure 7 shows that *Pernambuco* and *Beijing* are the best methods again. Since the performance of all methods (except for *Singapore*) decreases if evaluated on the ORAND-CAR-B dataset, it can be regarded as the most challenging in this contest. The precision of *Pernambuco* decreases to 0.754, while the precision of *Beijing* drops to 0.701.

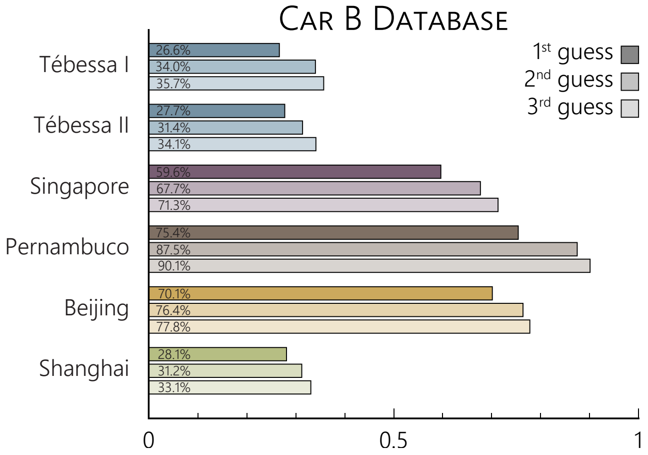

Figure 7. Precision of all participants on the CAR B database.

While both ORAND-CAR datasets show the performance on real world data with background clutter, the CVL dataset (see Figure 8) allows for empirically evaluating the methods using clean data.

A summary of the performance of each method on the three different databases is shown in Figure 9. Here, the dark bars represent the performance considering only the first guess, while the lighter bars represent the second and third guess respectively. Although, the CVL dataset contains clean digit strings, solely two participants (*Tébessa* and *Beijing* achieve better performance on this dataset. The methods from *Singapore* (-1.9%) and *Pernambuco* (-16.8%) achieve their worst performance on the CVL dataset if the first guess is considered. The former method achieves its

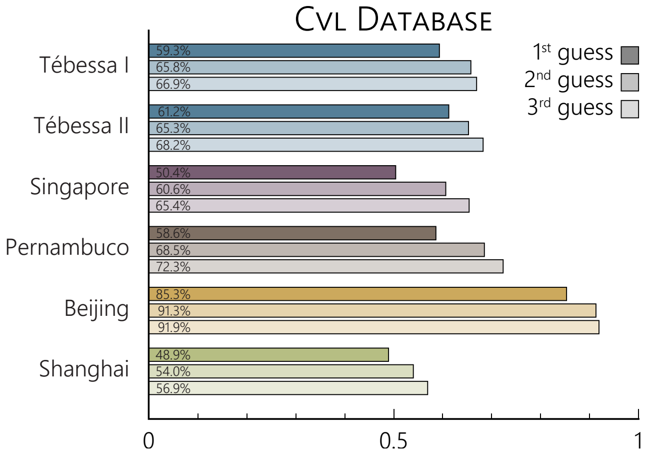

Figure 8. Precision of all participants on the CVL HDS database.

best results on the ORAND-CAR-B dataset which allows for drawing the conclusion that the approach is best at dealing with background clutter and noise. The approach submitted from *Pernambuco* has a significant performance drop if evaluated on clean data. This can be attributed to the fact that the system is designed for bank checks which are the basis of the ORAND-CAR datasets. Moreover, the removal of non-digit structures seems to degrade its performance is tested on clean data. Figure 9 additionally shows that the method submitted from *Singapore* and *Pernambuco* have the highest performance increase if the second guess is considered.

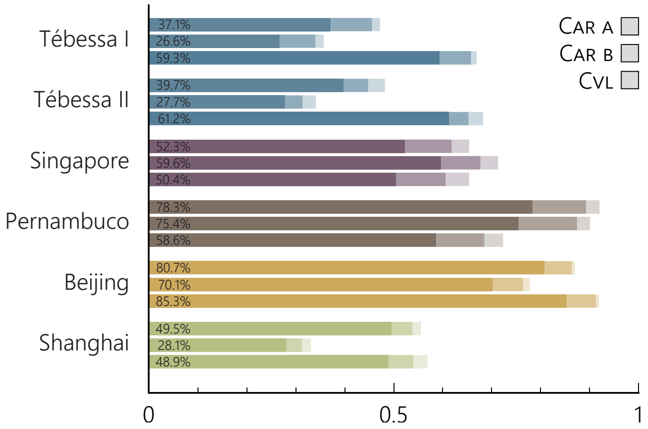

Figure 9. Precision of all participants on the three databases. The light bars represent the second guess and third guess respectively.

In Figure 10 the results achieved using the ANLD are shown. A lower value indicates a better performance. If the results of the ANLD are compared with those presented in Figure 9 are compared, it can be seen that *Shanghai* (28.1%) has a better performance than *Tébessa II* (27.7%) on the ORAND-CAR-B dataset. However, the error of the ANLD is larger for *Shanghai* (0.33 vs. 0.3) on this database. This indicates that *Tébessa II* classifies more digit strings false but with fewer errors (with respect to the edit distance).

It was previously mentioned, that *Singapore* achieves the worst performance on the CVL database. However, considering Figure 10, its ANLD error is lower (0.12) compared to the ORAND-CAR databases. This divergence results from two circumstances. First, the *Singapore* method has few errors per string and second, the average string length of the CVL database is larger (6.1) than that of ORAND-CAR-A (4.5) and ORAND-CAR-B (5.2).

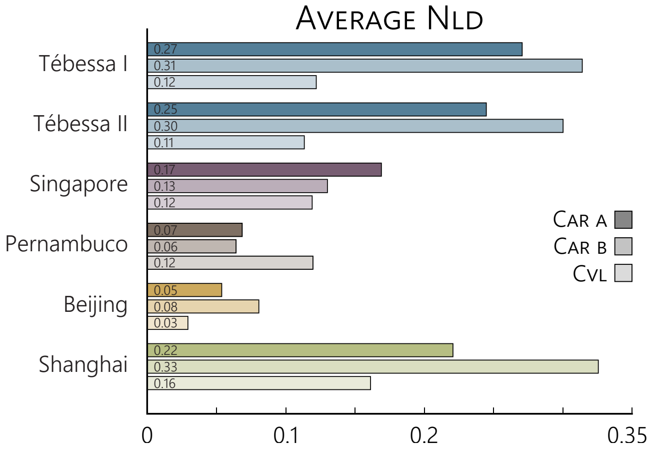

Figure 10. Average NLD for all participant methods.

Finally, in Figures 11, 12 and 13 we show a set of images for which the two best methods failed on the three different databases. Some factors that make a method fails are: out of focus, confusing strokes, and images with a low contrast.

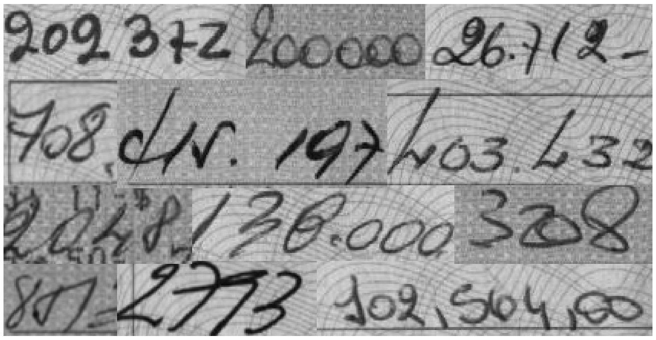

Figure 11. Images failed by the two best methods on the ORAND-CAR-A database.

V. CONCLUSION

In this paper we, have evaluated six recent approaches for handwriting digit string recognition. We presented two challenging databases, whose images represent digit strings of *real world* applications. The evaluation shows that the *Beijing* has the best recognition rate on two datasets. The *Pernambuco* method performs best if noisy images need to be recognized. In contrast to the other methods, it was

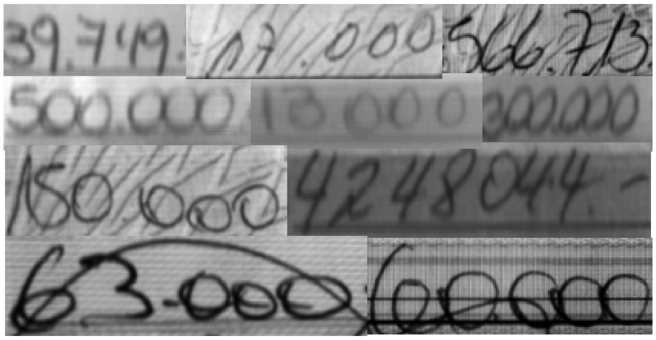

Figure 12. Images failed by the two best methods on the ORAND-CAR-B database.

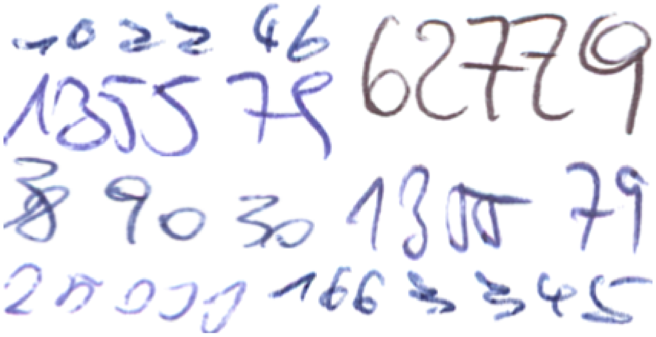

Figure 13. Images failed by the two best methods on the CVL HDS database.

designed for bank checks which are the basis for the ORAND-CAR databases. The *Singapore* and *Pernambuco* methods have the largest increase if the second guess is considered.

The empirical evaluation on the clean CVL database shows that digit string recognition is not yet solved if the results of the best performing method submitted from *Beijing* (85.3%) are regarded. Since all databases are publicly available[2], this competition can be used as baseline for benchmarking future approaches targeting digit string recognition.

Acknowledgment

We want to thank all authors who participated in this contest.

Additionally, the authors would like to thank to Juan Perez (BCI), Hugo Altamirano and Diego Muracciole (Precodata S.A.) for their worthy help to collect the ORAND-CAR images. Furthermore, we would like to thank all individuals who have contributed to the CVL Handwritten Digit database.

The research was funded by the Austrian Science Fund (FWF) project P23133 and CORFO-Chile project 12IDL-16293.

[2] http://www.orand.cl/en/icfhr2014-hdsr/

References

[1] M. Diem, S. Fiel, A. Garz, M. Keglevic, F. Kleber, and R. Sablatnig, “ICDAR 2013 Competition on Handwritten Digit Recognition (HDRC 2013),” in *Proc. of the 12th Int. Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 1454–1459.

[2] C.-L. Liu, H. Sako, and H. Fujisawa, “Effects of Classifier Structures and Training Regimes on Integrated Segmentation and Recognition of Handwritten Numeral Strings,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1395–1407, Nov. 2004.

[3] F. Yin, Q.-F. Wang, and C.-L. Liu, “Transcript Mapping for Handwritten Chinese Documents by Integrating Character Recognition Model and Geometric Context,” *Pattern Recogn.*, vol. 46, no. 10, pp. 2807–2818, Oct. 2013.

[4] Q.-F. Wang, F. Yin, and C.-L. Liu, “Integrating Language Model in Handwritten Chinese Text Recognition,” in *10th International Conference on Document Analysis and Recognition (ICDAR)*, July 2009, pp. 1036–1040.

[5] B. L. D. Bezerra, G. D. C. Cavalcanti, Z. C., and J. C. B. Rabelo, “Detecting and Treating Invasion in the Courtesy Amount Field on Bank Checks,” in *11th International Conference on Frontiers in Handwriting Recognition (ICFHR 2008)*, 2008.

[6] J. Rabelo, C. Zanchettin, C. A. B. Mello, and B. L. D. Bezerra, “A Multi-Layer Perceptron Approach to Threshold Documents With Complex Background,” in *2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2011, pp. 2523–2530.

[7] A. Graves and J. Schmidhuber, “Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks,” in *Advances in Neural Information Processing Systems 21*, 2008, pp. 545–552.

[8] B. L. D. Bezerra, C. Zanchettin, and V. B. de Andrade, “A MDRNN-SVM Hybrid Model for Cursive Offline Handwriting Recognition,” in *Proceedings of the 22nd International Conference on Artificial Neural Networks and Machine Learning - Volume Part II*, ser. ICANN’12, 2012, pp. 246–254.

[9] A. Gattal and Y. Chibani, “Segmentation Strategy of Handwritten Connected Digits (SSHCD),” in *Image Analysis and Processing – ICIAP 2011*, ser. Lecture Notes in Computer Science, G. Maino and G. Foresti, Eds., vol. 6979. Springer Berlin Heidelberg, 2011, pp. 248–254.

[10] ——, “Segmentation and Recognition Strategy of Handwritten Connected Digits Based on the Oriented Sliding Window,” in *2012 International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Sept 2012, pp. 297–301.