

# Knowledge Transfer for Writer Identification

Diego Bertolini<sup>1,2</sup>, Luiz S. Oliveira<sup>3</sup>,  
Yandre M. G. Costa<sup>2</sup>, and Lucas G. Helal<sup>2</sup>

<sup>1</sup>Federal University of Technology - Paraná (UTFPR), Campo Mourão, PR, Brazil

<sup>2</sup>State University of Maringá (UEM), Maringá, PR, Brazil

<sup>3</sup>Federal University of Paraná (UFPR), Curitiba, PR, Brazil

**Abstract.** The technical literature on writer identification usually considers the best case scenario in terms of data availability, i.e., a database composed of hundreds of writers with several documents per writer is available to train the machine learning models. However, in real-life problems such a database may not be available. In this context, learning from one dataset and transferring the knowledge to other would be extremely useful. In this paper we show how to transfer knowledge from one dataset to another through a framework that uses a writer-independent approach based on dissimilarity. Experiments on five different databases under single- and multi-script environments showed that the proposed approach achieves good results. This is an important contribution since it makes it possible to deploy the writer identification system even when no data from that particular writer are available for training.

## 1 Introduction

Writer identification is the task in which the goal is to identify whether the writer of a handwritten sample belongs to a set of writers or not. This matter has aroused interest once it can be used in several different applications, i.e. personalized handwriting recognition, writer retrieval, forensic document examination, and classification of ancient manuscripts. The literature shows that this problem has been attacked from several different perspectives for varied number of scripts, such as Latin [23,3,2], Arabic [6], Chinese [25], Japanese [14], among others. The amount of data used in terms of sample size may vary from several pages [3,5], paragraphs [1], to single words [2]. Recently, the writer identification problem was extended to a multi-script scenario where the rationale is recognizing an individual of a given text written in one script from the samples of the same individual written in another script [9,4].

A common trait among all the works reported in the literature is the dependence on the dataset. Usually, a dataset composed of  $n$  writers is divided into training and testing so that the writer is equally represented in both training and testing. The machine learning models are then trained on the training sets and results reported on the testing set. This approach is usually known as writer-dependent (WD). An alternative approach called writer-independent (WI) for writer identification was presented in [3], which converts the  $k$ -class

pattern recognition problem (where  $k$  is the number of writers) into a 2-class problem through a dichotomy transformation.

However, in both cases (WD and WI), writers used for training and testing are drawn from the same dataset, which in general contains a considerable number of individuals. Just to cite a few, IAM, IFN/ENIT, BFL, and CVL contains 650, 411, 315, and 310 writers, respectively.

In real-life application problem, though, such large datasets may not be available to train the machine learning models. A possible solution in this context is to learn in one dataset (source) and transfer the knowledge to other (target) dataset. This strategy has been explored in different domains of application with several different names, such as knowledge transfer, cumulative learning, life-long learning, transfer learning, etc. A review on this subject can be found in [21].

In this work we argue that the writer-independent (WI) approach based on dissimilarity described in [3] makes knowledge transfer possible because of its main property, i.e., the writer that did not contribute for the training set can still be identified by the system. To corroborate this hypothesis we carried out a series of experiments using five different datasets largely used in the literature (IAM [18], BFL [10], CVL [16], QUWI [17] and LAMIS [8]) and two different textural descriptors, the Local Binary Pattern (LBP)[19] and Robust Local Binary Pattern (RLBP) [7]. The selected databases contain documents written in different languages (English, French, German, Portuguese, and Arabic) and two different scripts (Roman and Arabic).

In the first experiment we considered only Roman scripts where two databases were used as source and other three as target. In a second experiment we evaluated if the results found for Roman script can be replicated for Arabic scripts. Finally, in the third experiment, we addressed the multi-script problem.

## 2 Knowledge Transfer with Dissimilarity

As mentioned before, an interesting aspect about the dissimilarity approach is the possibility of reducing any pattern recognition problem to a 2-class problem. The idea consists in extracting the feature vectors from both questioned and reference samples and then computing what we call the dissimilarity feature vectors. In ideal conditions, it is expected that if both samples come from the same writer (genuine), then all the components of such a vector should be close to 0, otherwise, the components should be far from 0.

Given a queried handwritten document and a reference handwritten document, the goal consists in determining whether or not the two documents were produced by the same writer. Let  $V$  and  $Q$  be two vectors in the feature space, labeled  $l_V$  and  $l_Q$  respectively. Let  $Z$  be the dissimilarity feature vector resulting from the dichotomy transformation  $Z = |V - Q|$ , where  $|\cdot|$  is the absolute value. This dissimilarity feature vector has the same dimensionality as  $V$  and  $Q$ .

In the dissimilarity space, there are two classes that are independent of the number of writers: the within class (+) and the between class (o). The dissimilarity vector  $Z$  is assigned the label  $l_Z$ ,

$$l_Z = \begin{cases} + & \text{if } l_V = l_Q, \\ o & \text{otherwise} \end{cases} \quad (1)$$

The rationale of knowledge transfer using the dissimilarity approach is presented in Figure 1. In this example, (a) depicts the source dataset in the feature space from five different writers while (b) shows the dissimilarity vectors for the source dataset, which are the results of the dichotomy transformation between the features of each pair of samples to form vectors. The same representation for the target dataset is presented in Figures 1(c) and (d). Differently of the feature space, where multiple boundaries are necessary to discriminate the writers, in the dissimilarity space, only one boundary is necessary, since the problem is reduced to a 2-class classification problem. Besides, even writers whose specimens were not used for training can be identified by the system. This characteristic is quite attractive, since it obviates the need to train a new model every time a new writer is introduced.

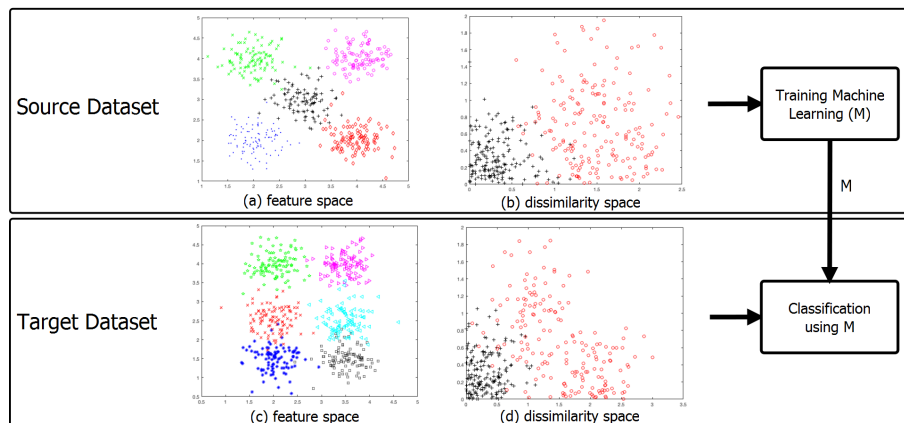


Fig. 1: Knowledge Transfer with Dissimilarity: (a) samples of the source dataset in the feature space, (b) samples of the source dataset in the dissimilarity space, (c) samples of the target dataset in the feature space, and (d) samples of the target dataset in the dissimilarity space. In (b) and (d), “+” stands for the vectors associated to the within class and “o” stands for the vectors associated to the between class..

In spite of the different number of writers and distributions of the feature space observed in the source and target datasets (Figures 1(a) and (c)), the dichotomy transformation impacts in the geometry producing very similar distribution for both source and target datasets in the dissimilarity space (Figures 1(b) and (d)). This observation leads us to investigate whether knowledge transfer is an alternative in this context, i.e., to learn the machine learning model ( $M$ ) in the source dissimilarity space and deploying it to other target dataset.

In order to generate the positive samples (+) to train the classifier, we computed the dissimilarity vectors among the  $R$  genuine samples (references) of each writer which resulted in  $\binom{R}{2}$  different combinations. The same number of

negative samples (o) is generated by computing the dissimilarity between one reference of one writer against one reference of other writers picked at random. In our experiments, the best results were found using 9 references per writer.

### 3 Databases and Feature Extraction

As mentioned earlier, we have selected five different databases to be used as source and target in our experiments. These databases contain handwriting in different languages and scripts.

The IAM dataset [18] is widely used in problems such as handwriting recognition and writer identification. It contains forms with handwritten English text of variable content (text-independent). A total of 650 writers have contributed to the dataset. BFL database [10] is composed of 315 writers, with three samples per writer, for a total of 945 images. This makes it suitable for text-dependent writer identification as well. In the CVL database [16] contains 310 writers with 1,604 text-dependent handwriting. Furthermore, 27 writers have six documents in English and one in German, while 283 writers have four documents in English and one in German. The LAMIS [8] contains 1,200 handwritten text images from 100 different writers. To acquire the handwritten samples writers were instructed to produce 12 pages of handwriting, six of them in French and other six in Arabic, all the letters are text dependent. To the best of our knowledge, this database has not been used for writer identification. Lastly, the QUWI database [17] contains 4,068 handwritten text images from 1,017 different writers. In order to acquire the handwritten samples, volunteers were instructed to produce four pages of handwriting as follows: First and second pages in the Arabic language free-text and copied or (text-independent). The third and fourth page in English of free-text and copied, respectively.

The recent literature on writer identification shows that researchers have been investigating a great number of representations for writer identification, which can be classified into local and global. The local approach takes into account specific features of the writing and generally involve some kind of segmentation process. Such features are usually extracted from words, characters, or allographs [2,22]. The global approach, on the other hand, tries to avoid the segmentation process by representing the handwriting as a texture [12,3,11]. In these cases, several textural descriptors have been tried out such as Grey-Level Co-occurrence Matrices (GLCM) [13], Local Phase Quantization (LPQ)[20], Local Binary Patterns (LBP) [19] and its variations such as Robust LBP [7].

In this work we adopted the global approach described by Bertolini et al. in [3]. In this case, a texture of the handwriting is created scanning the document top-down left-write and putting together all the connected components found in the image. Small components, such as strokes, commas, and noise are discarded. Then, the texture is segmented into nine  $256 \times 256$  blocks. Figure 2 shows two examples of the handwriting texture produced from English and Arabic handwritings for the same writer of the LAMIS database.

As representation, we have tested several textural descriptors but our best results were always achieved either with LBP or RLBP. Both descriptors are

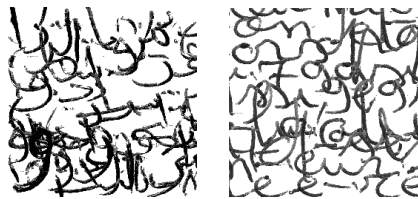


Fig. 2: Example of the texture produced from the LAMIS Database (a) Arabic and (b) English handwritings for the same writer.

quite similar, but RLBP considers a more flexible concept of uniformity. For example, if there is one, and only one value in the binary code which makes the LBP non-uniform, it is possibly caused by some noise and, for this reason, it must be considered as a uniform pattern. In our experiments, both LBP and RLBP representations produce a 59-dimensional feature vector. In both cases, the feature vectors are normalized with the min-max rule. For those readers interested in these textural descriptors, please refer to [19] and [7].

#### 4 Experimental results and discussion

In all experiments described in this work, the Support Vector Machine (SVM) was used to perform the classification [24]. The parameters of the system for training were chosen using 5-fold cross validation. The best results were achieved using Gaussian kernel. Parameters  $C$  and  $\gamma$  were determined through a grid search.

To corroborate the hypothesis presented in the introduction we have designed a set of three experiments. In the first experiment we considered only Roman scripts where two databases were used as source (QUWI and LAMIS) and other three as target (IAM, BFL, CVL). To better compare the results, we fixed in 100 the number of writers for both source datasets, which were picked randomly.

Table 1 summarizes the first experiment. In the first part of the table we report the performance on the three target datasets using QUWI, LAMIS, and the union of both as source data. The second part of the table allows us to better assess these results since source and target are the same. As we can observe, the results using knowledge transfer are quite similar to those where the same dataset was used for training and testing. In the case of IAM, using QUWI or LAMIS as source produced better results and using IAM itself. Besides, we notice that the source data can benefit from a larger number of writers. When both source datasets were combined (LAMIS+QUWI), we perceived some improvement in all three target datasets. Finally, the best results were achieved in the BFL dataset, which is a text-dependent database that contains a great deal of handwriting so that we have no problems in generating the same amount of texture for all writers. In the case of CVL and IAM, some writers are represented by just one letter, hence, compromising the texture generation process. Table 1 also shows the good performance of the knowledge transfer when source and target contain handwriting in different languages, using the Roman script, though. In this case, the source datasets contain handwriting in English (QUWI) and French (LAMIS)

Table 1: Performance on IAM, BFL, and CVL using the models trained on QUWI and LAMIS. The number of writers used is in parenthesis

Source	Target					
	BFL (315)		CVL (310)		IAM (650)	
	LBP	RLBP	LBP	RLBP	LPB	RLBP
QUWI (100)	97.8	99.3	91.6	93.8	93.5	92.9
LAMIS(100)	98.4	99.4	89.7	91.9	92.3	92.5
QUWI+LAMIS (200)	99.0	99.6	93.5	92.9	94.8	93.3
BFL (100)	99.0	99.0				
CVL (100)			93.5	93.8		
IAM (100)					89.0	90.0

while the target datasets are written in Portuguese (BFL), English/German (CVL). Before discussing the second experiment, it is important mentioning that the literature shows identification rates on IAM and CVL around 97% and 99%, respectively [15]. In these cases, the same datasets are used as source and target and all the same users are considered for both training and testing.

So far all the experiments have considered only the Roman script. Since both QUWI and LAMIS contain documents written in Arabic, we assessed the knowledge transfer applied for Arabic. It is noteworthy that the QUWI database contains two types of documents, free and copy text. The free-text is composed of approximately six lines of handwriting in Arabic while the copy text contains three paragraphs of Arabic handwriting. In other words, the copy text has considerably more handwriting, which translates directly in a richer texture. In this experiment, 100 writers were considered in both source and target datasets. In the case of the QUWI, only copy documents were used.

Table 2 reports the results of the knowledge transfer for the Arabic script. Similarly to the experiments on the Roman script, the knowledge transfer achieves good performance in both cases. The discrepancy in terms of performance between both copy and free texts is related to the different amount of text available for copy and free text. Since the samples of free texts are composed of few lines, we could not create dense textures as we did on copied texts, which are composed of three paragraphs.

Table 2: Performance of the knowledge transfer for Arabic scripts.

Source	Target					
	LAMIS		QUWI-Free		QUWI-Copy	
	LBP	RLBP	LBP	RLBP	LPB	RLBP
QUWI	97.0	94.0	77.0	79.0	97.0	97.0
LAMIS	94.0	97.0	72.0	69.0	98.0	99.0

Finally, we address the problem of multi-script where the idea is to identify an individual of a given text written in one script from the samples of the same individual written in another script. In the context of knowledge transfer, the source dataset could be Arabic and the target in Roman, and vice-versa.

Similarly to the previous experiments, 100 writers were considered in both source and target datasets. Table 3 presents the performance of the knowledge transfer for multi-script where LAMIS and QUWI were used as source and target. In the first part of Table 3 the model was trained on Roman while the testing was performed on Arabic. The second part of the table the scripts were inverted.

Table 3: Performance of the knowledge transfer for multi-script.

Script Train/Script Test	Source	Target	LBP	RLBP
Roman/Arabic	LAMIS	QUWI-free	72.0	74.0
	LAMIS	QUWI-copy	100.0	98.0
	QUWI-copy	LAMIS	98.0	93.0
	QUWI-free	LAMIS	97.0	98.0
Arabic/Roman	LAMIS	QUWI-free	75.0	83.0
	LAMIS	QUWI-copy	97.0	98.0
	QUWI-copy	LAMIS	98.0	98.0
	QUWI-free	LAMIS	96.0	99.0

The results presented in Table 3 show the robustness of this approach. In all tests we have performed (except when QUWI-free has been used as target, as explained before) the results on multi-script compare to single-script results. This shows that this strategy is viable for both these cases.

## 5 Conclusion

In this work we have shown through a series of experiments on five different databases using two different textural descriptors that the writer-independent approach underpinned on the concept of dissimilarity allows knowledge transfer for writer identification. We have assessed the proposed approach single-script (Roman/Arabic) and multi-script environments and observed that in all cases one can transfer the knowledge from one dataset to another. This is an important contribution since it makes it possible to deploy the writer identification system even when no data from that particular writer are available for training.

## References

1. S. Al-Maadeed. ext-dependent writer identification for arabic handwriting. *Journal of Electrical Comput. Eng.*, 2012.
2. A. Bensefia and T. Paquet. Writer identification based on a single handwriting word samples. *EURASIP Journal on Image And Video Processing*, 2016.
3. D. Bertolini, L. S. Oliveira, E. Justino, and R. Sabourin. Texture-based descriptors for writer identification and verification. *Expert Systems With Applications*, 40(6):2069–2080, 2013.
4. D. Bertolini, L. S. Oliveira, and R. Sabourin. Multi-script writer identification using dissimilarity. In *International Conference on Pattern Recognition (ICPR)*, pages 3020–3025. IAPR, 2016.
5. A. Brink, L. Schomaker, and M. Bulacu. Towards explainable writer verification and identification using vantage writers. In *9th International Conference on Document Analysis and Recognition*, pages 824–828, Curitiba, Brazil, 2007.
6. Marius Bulacu and Lambert Schomaker. Text-independent writer identification and verification using textural and allographic features. *IEEE transactions on pattern analysis and machine intelligence*, 29(4), 2007.

7. J. Chen, V. Kellokumpu, G. Zhao, and M. Pietikäinen. RLBP: Robust Local Binary Pattern. In *Proceedings of the British machine vision conference*, 2013.
8. C. Djeddi, A. Gattal, L. Souici-Meslat, I. Siddiqi, Y. Chibani, and H. Abed. Lamismshd: A multi-script offline handwriting database. In *ICFHR*, pages 93–97, 2014.
9. C. Djeddi, I. Siddiqi, L. Souici-Meslati, and A. Ennaji. Text-independent writer recognition using multi-script handwritten texts. *Pattern Recognition Letters*, 34:1196–1202, 2013.
10. C. Freitas, Luiz S Oliveira, Robert Sabourin, and Flávio Bortolozzi. Brazilian forensic letter database. In *11th International workshop on frontiers on handwriting recognition, Montreal, Canada*, 2008.
11. Y. Hannad, I. Siddiqi, and M. Kettani. Writer identification using texture descriptors of handwritten fragments. *Expert Systems with Apps.*, 47:14–22, 2016.
12. Regiane Kowalek Hanusiak, Luiz S Oliveira, E Justino, and Robert Sabourin. Writer verification using texture-based features. *International Journal on Document Analysis and Recognition (IJDAR)*, 15(3):213–226, 2012.
13. R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Trans. on systems, man, and cybernetics*, 3(6):610–621, 1973.
14. Hiroshi Kameya, Shunji Mori, and Ryuichi Oka. A segmentation-free biometric writer verification method based on continuous dynamic programming. *Pattern Recognition Letters*, 27(6):567–577, 2006.
15. F. A. Khan, M. Tahir, F. Khelifi, A. Bouridane, and R. Almotaeryi. Robust off-line text independent writer identification using bagged discrete cosine transform features. *Expert Systems with Applications*, 71:404 – 415, 2017.
16. F. Kleber, S. Fiel, M. Diem, and R. Sablatnig. Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In *12th International Conference on Document Analysis and Recognition*, pages 560–564. IEEE, 2013.
17. S. A. Maadeed, W. Ayouby, A. Hassaïne, and J. M. Aljaam. Quwi: An arabic and english handwriting dataset for offline writer identification. In *2012 International Conference on Frontiers in Handwriting Recognition*, pages 746–751, Sept 2012.
18. U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46, 2002.
19. Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.
20. Ville Ojansivu and Janne Heikkilä. Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing*, pages 236–243. Springer, 2008.
21. S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
22. I. Siddiqi, C. Djeddi, A. Raza, and L. Souici-meslati. Automatic analysis of handwriting for gender classification. *Pattern Analysis and Apps.*, 18(4):887–899, 2015.
23. Imran Siddiqi and Nicole Vincent. Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. *Pattern Recognition*, 43(11):3853–3865, 2010.
24. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, NY, 1995.
25. Xianliang Wang and Xiaoqing Ding. An effective writer verification algorithm using negative samples. In *Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on*, pages 509–513. IEEE, 2004.