Towards Visuo-Structural Handwriting Evaluation Based on Graph Matching

Chapter · October 2023

DOI: 10.1007/978-3-031-45461-5_6

5 authors, including:

Céline Rémi
Université des Antilles

Emmanuel Biabiany
Université des Antilles

Jimmy Nagau
Université des Antilles

Towards Visuo-Structural Handwriting Evaluation Based on Graph Matching

Anna Scius-Bertrand[1,2], Céline Rémi[3(✉)], Emmanuel Biabiany[3], Jimmy Nagau[3], and Andreas Fischer[1,2]

[1] University of Applied Sciences and Arts Western Switzerland, Fribourg, Switzerland
{anna.scius-bertrand,andreas.fischer}@hefr.ch
[2] University of Fribourg, Fribourg, Switzerland
[3] Université des Antilles, B.P. 592, 97157 Pointe à Pitre Cedex, France
{celine.remi,emmanuel.biabiany,jimmy.nagau}@univ-antilles.fr

Abstract. Judging the quality of handwriting based on visuo-structural criteria is fundamental for teachers when accompanying children who are learning to write. Automatic methods for quality assessment can support teachers when dealing with a large number of handwritings, in order to identify children who are having difficulties. In this paper, we investigate the potential of graph-based handwriting representation and graph matching to capture visuo-structural features and determine the legibility of cursive handwriting. On a comprehensive dataset of words written by children aged from 3 to 11 years, we compare the judgment of human experts with a graph-based analysis, both with respect to classification and clustering. The results are promising and highlight the potential of graph-based methods for handwriting evaluation.

Keywords: scholar handwriting · legibility · children · graph-matching · similarity metrics · clustering

1 Introduction

Handwriting remains a skill and a crucial mode of communication in our societies. As a result, the characterization of the quality of handwritten traces is a problem shared by multiple communities of researchers and experts who must manipulate and process such offline and online productions [13]. The literature reports numerous works on evaluation of the quality of written productions [23]. Then the specialists of reeducation of non-proficient handwriting, as well as researchers interested in the acquisition of handwriting, have various batteries of tests for the analytical evaluation of the quality of children [8,12,14,19] as well as adolescents [15,28] handwriting. Whether the methods for such qualitative evaluation are global or analytical, they are based on visuo-structural criteria. Those criteria are estimated on specific sequences of patterns of handwriting, like word, sentence, paragraph or text, thanks to psychometric exercises. These criteria have usually to reflect both: the sharpness and fidelity of the form of

A. Parziale et al. (Eds.): IGS 2023, LNCS 14285, pp. 75–88, 2023.
https://doi.org/10.1007/978-3-031-45461-5_6

each of the symbols or trajectories constituting the handwritten word produced in relation to some reference models taught at school; the spatial organization of their traces and their proportions within the plot space considering the writing conventions taught; or their spatial organization and proportions in relation to each other.

So, considering those visuo-structural criteria seems fundamental for human experts when analyzing the quality of children's handwriting. It is also an important issue for the processes of comprehension [22] and accompaniment [11,30] of learning to write. This, whether these processes take place in schools and whether they involve the use of digital applications or not. Whatever the education system, official school curricula clearly stipulate the importance for the pupil of acquiring and then maintaining the criterion of legibility alongside fluidity [1]. This, throughout the school curriculum that will see him go from the status of novice apprentice scripter to that of adolescent learner mastering the gestures and principles of written expression. Also, in the case of the school context, the evaluation of the visuo-structural quality carried out by the expert teacher consists in a visual global judgement of the legibility of the written productions of the pupils. However, more often, the global criterion of legibility turns out to be partly subjective since teachers have neither common training in evaluation of this criterion, nor shared principles or tools for evaluating it so that it is otherwise [2]. Some previous works have been realized to deal with such problem and to develop new tools for class teachers like the Handwriting Legibility Scale (HLS) [3] for a more objective global scoring of legibility at school. However, we assume that computerized methods would be of great help to provide a uniform basis for evaluation of legibility and to deal with large quantities of handwriting samples, such that children with difficulties can be identified and accompanied by the teacher.

This observation and many others, led the University of the West Indies and the Regional Academy of Guadeloupe to initiate together a project, propelled by the application Copilotr@ce [20]. This project focuses for the moment on the French language that prevails in the teachings provided in their territories of establishment. It aims to make local teaching teams of the first and second degree collaborate with researchers for the development of collaborative digital tools of assistance: individualized and continuous support of learning from kindergarten to entry into college, help in identifying and remediating identified difficulties. In the case of learning to write, one of the challenges is the design of objective and automated solutions for the assessment of readability. These solutions must behave in terms of judgment that are as consistent and faithful as possible to that of a pool of expert teachers confronted with the same set of plots.

Among the possible solutions to this end, this article proposes to explore the use of Graph-based methods from structural pattern recognition [6]. We have chosen to firstly explore these methods because they are promising ways to capture and analyze the global structure of the handwriting based on images, as it has been put in evidence by studies on handwriting recognition [9], handwritten keyword spotting [21,27,29], and signature verification [18], to name just a few.

In the present paper, we investigate the potential of graph-based methods for the automatic qualitative evaluation of isolated cursive words handwritten by students. Our study has two main objectives, which we will refer to as O1 and O2.

O1- First, it is necessary to establish whether usual measure of similarity of graphs constructed from visuo-structural data extracted from offline images of handwritten words can allow decision-making analogous to those of human experts concerning:

– the similarity of the visuo-structural quality of students' handwritten productions?
– the legibility of handwritten productions of words?

O2- Next, we need to assess whether a scoring based on such measures of similarity could contribute to the identification of relevant handwritten words groupings, i.e., models meaning from a school point of view.

To test and to compare the results of our graph-based method, we needed some criteria and ground truth annotations. So, the first challenge for qualitative evaluation consists in choosing criteria that match with usual human scholar features used for qualitative evaluations. We have chosen to retain the overall legibility of the handwritten words. Therefore, results provided by our automated graph-based approach for objectives O1-a and O2 will be confronted with the qualitative criterion of legibility for both the O1 and O2 objectives. This article will be structured as follows. First, the next section will develop the context and the methodology used to produce the dataset and its ground truth. Next, the structural graph-based method we have chosen will be introduced and described. Subsequently, the experimental results will be presented and discussed. Finally, we draw some conclusions.

2 Dataset Used for This Study

2.1 Tool Used for Online Acquisition and Human Evaluation of Images

Trace acquisition and evaluation were carried out using the Copilotr@ce [20] application. Copilotr@ce is a web application for capturing handwritten gestures, available on all types of screen-based hardware.

Copilotr@ce works with or without an Internet connection, in online or offline mode. It offers the possibility of directly displaying or replaying traces made on its platform during acquisition campaigns.

The handwritten gesture capture session can be contextualized by activities requiring the use of graphomotor gestures.

Copilotr@ce captures and records over time sequences of points produced by the movement of a writing tool on the work surface at a frequency of 100 Hz. Depending on the hardware, these sequences can be produced using a finger, a writing tool such as a stylus for touch-screen hardware, or a mouse for computers with traditional screens.

Depending on the configuration chosen for an experiment, it is possible to start recording a sequence of points at the start of an activity, or from the first contact between the writing tool and the work surface.

In addition to the raw trace data recorded in real time (dating, coordinates, pressure, etc.), Copilotr@ce provides a set of indicators derived from the trace during and at the end of the experiment. These data can be used in models to provide evaluation and positioning indicators.

Copilotr@ce enables the collection of contextualized traces as part of action research. These collections are programmed on cohorts of anonymized scribblers.

Copilotr@ce enables the evaluation of traces contained in its information database, depending on the study context, by cohorts of evaluators or experts. The information databases contained in Copilotr@ce are represented by: activity contexts, writers' traces, as well as human or automatic evaluations. All this data is used to conduct studies with the aim of evaluating and building automatic analysis models, which can then be fed back into its knowledge base for validation in the field.

2.2 Nature and Context of the Image Acquisition

The partnership project powered by the Copilotr@ce platform, has already allowed to collect a substantial mass of handwritten traces of great diversity produced by pupils from 3 up to 14 years old. Among those handwritten traces we have chosen to consider 814 images of handwritten isolated words collected for this study. They were handwritten by 321 all-comers aged 3 to 11 years old from kindergarten to middle school. All these traces were made by these pupils with a stylus on touch tablets during a task of copying cursive models of each of the words: "lundi" (Monday), "lunes" (moons), and "plumes" (feathers) (see Figs. 1 and 2). We have chosen these three words because when we began this present study, they were those that were more represented in the dataset created thanks to Copilotr@ce. Indeed, they had been the most frequently and spontaneously handwritten among all those that were presented to the participants in the scholar action-researches driven from the first grade of kindergarten up to the first one of the middle school.

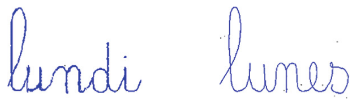

Fig. 1. Cursive patterns of the words "lundi" and "lunes" presented during the copy task.

This copy task was proposed by the Copilotr@ce application to pupils during school time according to the same modalities. As the successive presentations of the isolated words on the screen, the students copied them into a reserved area with their finger or stylus on the surface of the touch tablet. This area could present a baseline as shown by Fig. 2 for the word "plumes". Figure 3 shows three examples of productions for pupils who participated in this activity.

Fig. 2. The model of the word "plumes" and baseline presented by Copilotr@ce during the copy task.

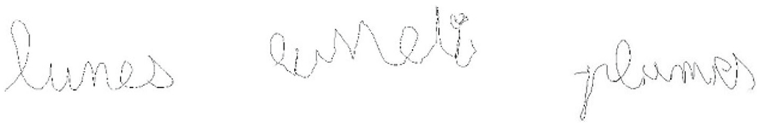

Fig. 3. Examples of copies of the words "lunes", "lundi" and "plumes" made by 3 students.

2.3 Description of the Ground Truth

The ground truth was built by mobilizing three of the co-authors of this contribution who are also teachers. The latter took no part in the @MaGma project either as teachers in one of the participating classes or as accompanists of cohorts of pupils that had handwritten the words which are considered in this study.

We provide two levels of ground truth for each handwritten word: one with two classes "legible" and "illegible" and one with three classes "legible", "not very legible", and "illegible". First, the two-class annotation is performed and then, in a second step, for some of the samples the third class "not very legible" is attributed. This third class represents uncertainty of the human experts and concerns both samples previously labeled as legible and illegible. The human experts did not agree among themselves on all samples. We use majority voting to assign a final label to each handwritten word.

3 Methods

To classify the children's handwriting, we use graph matching. Once the similarity between each graph has been calculated, we compare two methods to assign the closest class: classification with KNN and clustering with K-Medoids and Agglomerative Hierarchical Clustering.

3.1 Graph-Based Approach Principle

A graph is a mathematical representation of the components of an object and the relationships between them, such as molecules with linked atoms, proteins with linked amino acids – or handwriting with linked strokes. It is called a structural representation because it captures the global structure of the object. Representing handwriting by graphs enables us, among other things, to compare the similarity between two words. We assume that when comparing a set of handwritten words, words categorized as legible should be the most similar to each other; and the same for not very legible and illegible words. This would enable us to identify students in need of remediation.

Graph Definition. A graph g is defined by four components:

$$g = (V, E, \mu, \nu) \tag{1}$$

where V is a finite set of nodes, E a set of edges with $E \subseteq V \times V$, $\mu : V \rightarrow L$ corresponds to the labels of the nodes and $\nu : E \rightarrow L$ corresponds to the labels of the edges.

A graph may or may not have labels and may or may not be directed. Graphs whose edges have no direction are undirected graphs. Conversely, graphs whose edges have a direction are called directed graphs. Nodes and/or edges can have labels. Labels can be part of any domain, they can be numerical (L = 1, 2, ..., n) or vectorial (L = $\mathbb{R}^n$) or symbolic (L = $\{\alpha, \beta, ..., n\}$) or even a set of colors (L = {violet, yellow, green, ...}).

Graph Extraction. The first step in comparing two graphs is to extract graphs from each of the word images to be matched. We have chosen keypoint graphs [9] as our graph representation, as they allow us to represent the trace of a word as closely as possible. Furthermore, they have shown very good results in handwriting analysis [18,27,29].

Formally, keypoint graphs use coordinates $(x, y) \in \mathbb{R}^2$ as node labels and edges are unlabelled. Note that a relatively large number of nodes replaces the need for more complex edge labels. Adding edge labels, such as distances or angles, have not led to improved performance for handwriting analysis in preliminary experiments.

To extract keypoint graphs, first, a difference of Gaussians filter (DoG) is applied to enhance the edges. Next, a binarization is performed with a global threshold. Then a skeleton is extracted by reducing the thickness of each line to one pixel. Three types of points are then detected: stroke ends, intersections and a random point on circular structures. We then add additional points on the skeleton at distance D. Each point becomes a node, and the strokes between each point become edges. A visual representation of a word sample is provided Fig. 4.

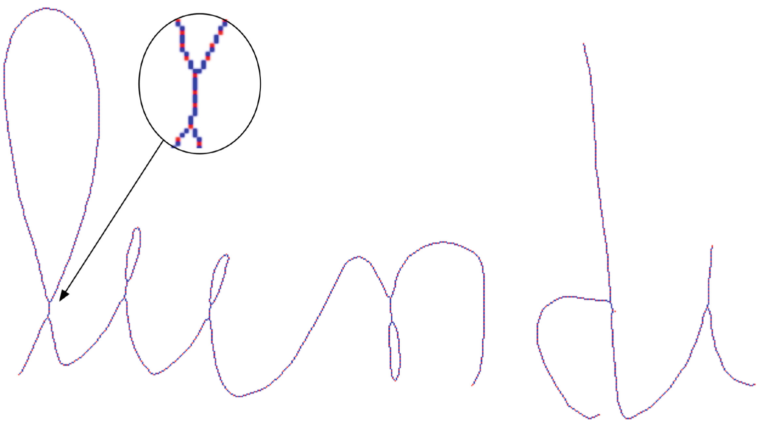

Fig. 4. Graph representation of a Monday's sample. Nodes are in red and edge are in blue. A closer look is made for the intersection strokes of the l. (Color figure online)

Graph Matching. Once we have obtained a set of graphs, Graph Edit Distance (GED) allows us to calculate a minimal transformation cost between two graphs. The cost takes into account to node deletion ($u \rightarrow \epsilon$), node insertion ($\epsilon \rightarrow v$), node label substitution ($u \rightarrow v$), edge deletion ($s \rightarrow \epsilon$), and edge insertion ($\epsilon \rightarrow t$). However, GED is NP-complete, which makes the computation infeasible when a graph has more than a few dozen nodes. This is why we use the Hausdorff Edit Distance (HED) [10] to compute a lower bound approximation in quadratic time:

$$HED_c(g_1, g_2) = \sum_{u \in V_1} \min_{v \in V_2 \cup \{\epsilon\}} f_c(u, v) + \sum_{v \in V_2} \min_{u \in V_1 \cup \{\epsilon\}} f_c(u, v) \tag{2}$$

where c is the cost function for the edit operations and $f_c(u, v)$ the cost for assigning node u to node v, taking into account its adjacent edges as well.

The Euclidean cost function is used, i.e. constant costs c_V and c_E

$$\begin{aligned} c(u \rightarrow \epsilon) = c(\epsilon \rightarrow v) = c_V \\ c(s \rightarrow \epsilon) = c(\epsilon \rightarrow t) = c_E \end{aligned} \tag{3}$$

for node and edge deletion and insertion, and the Euclidean distance

$$c(u \rightarrow v) = ||(x_u, y_u) - (x_v, y_v)|| \tag{4}$$

for node label substitution.

3.2 Classification of Graphs Using Similarity Measures

The first objective of our study (O1) is concerned with comparing automatic classification with human judgment. For this prupose, we use a standard classifier that operates directly on the pairwise dissimilarity obtained by HED, namely k-nearest neighbors (KNN) classification. It compares a test graph with a set of training graph and selects the k most similar training samples with respect to the dissimilarity measure, HED in our case. Afterwards the class that is most frequent among the k nearest neighbors is chosen as the class of the test graph. In the case of a tie, the class of the nearest neighbor is chosen.

We are using a simple accuracy measure (see Eq. 5) to evaluate the performance regarding the classification of the samples. We count for the whole test set the amount of correctly classified samples and divide it by the size of the test set.

$$Accuracy = \frac{\#correct}{\#total} \tag{5}$$

The results of this classification approach are presented and discussed in a Subsect. 4.2.

3.3 Clustering of Graphs Using Similarity Measures

The second objective of our study (O2) is concerned with assessing the graph-based similarity measure with respect to its ability to group handwritten words that share visuo-structural features. For this purpose, we consider clustering algorithms.

We first determine the dissimilarities between all graphs using HED, thereby producing a distance matrix. Secondly, this matrix is used by clustering algorithms as custom metric in order to identify groups of homogeneous graphs (with the same characteristics according to the similarity measure). The use of a metric adjusted to the specific data and constraints of a study makes it possible to obtain relevant results in the field of climate informatics [4,5], and this experiment aims to verify this principle on another field of research.

K-Medoïds (KMED) and Agglomerative Hierarchical Clustering (AHC) algorithms are used with different settings in order to compare their results [16,17, 25,26]. The number of clusters (k), the choice of algorithm and the quality of the grouping will be determined using the Silhouette index [7,24]. This index varies between −1 and 1, with negative values indicating the absence of data patterning, a value of 0 indicating the presence of a single group and values above 0.2 indicating the presence of data patterning. Therefore, the higher the index, the more relevant the clustering.

The results of this clustering approach are presented and discussed in a Subsect. 4.3.

4 Experimental Evaluation

To evaluate our method for classifying the legibility of handwritten words by children, we conducted a series of experiments. First, we optimized the parameters of the graphs on the words at our disposal. Then we calculated the distance between words. Next, we interpreted these distances from the perspective of classification and clustering, respectively.

In the following, we first describe our experimental setup, then the results with classification and clustering.

4.1 Parameter Optimisation

In order to benefit from as much data as possible to optimize parameters and test our method, we have divided our dataset into two parts: validation set (30%) and testset (70%). Tables 1 and 2 list the number of words in the test set for the three and two classes, respectively.

Table 1. Test set of handwritten words with three classes.

	Lundi	Lunes	Plumes
Legible	141	183	15
Not very legible	12	54	49
Illegible	21	63	35
Total	174	300	99

Table 2. Test set of handwritten words with two classes.

	Lundi	Lunes	Plumes
Legible	141	230	63
Illegible	33	70	36
Total	174	300	99

The parameter optimisation is performed with respect to KNN-based classification on the validation set using a leave-one-out strategy, i.e. each sample of the validation set is classified with respect to all others. The setup is as follows. Node labels have been normalized to a zero mean and a unit variance (z-score), since word positions vary significantly from one example to another (top left, center, bottom...). For the optimization, we evaluated several parameters:

- **For graph extraction**, we tested different node distance values: $D \in \{3, 5, 10, 15\}$.
- **For graph matching parameters**, we tested the following values for node costs c_V and edge costs c_E: $c_V, c_E \in \{0.5, 1.0, 1.5\}$.
- **For classification**, we tested the following values of k for KNN-based classification: $k \in \{1, 3, 5, 7, 9\}$.

The optimization result for each word is shown in Table 3. The first parameter set P1, which has been optimized for the word “lundi”, favors graphs with a high resolution (small node distance D = 5 on the skeleton) and has relatively low costs for node insertion/deletion (0.5 standard deviations). P2 and P3 have a lower resolution and higher node/edge costs.

Table 3. Meta-parameters after optimization.

Parameters	P1 (lundi)	P2 (lunes)	P3 (plumes)
D	5	10	10
Node cost c_V	0.5	1.0	1.5
Edge cost c_E	1.0	1.0	1.5
k	1	5	5

4.2 Results with Classification

To evaluate the automatic evaluation of handwriting, we perform a KNN-based classification on the test set. Tables 4 and 5 show the accuracy results obtained for three and two classes, respectively.

For three classes (legible, not very legible, illegible), we achieve a promising accuracy between 68.7–82.2%. For two classes (legible, illegible) the accuracy is even better, between 78.8–85.0% depending on the word. The improved accuracy for two classes is as expected, because the classification task is simplified, focusing only on the two extreme cases.

In all cases, the parameter set P2 is the best, which was optimized for the word "lunes" during validation. A possible explanation is that this word had the largest number of samples in the validation set, which leads to a more stable estimation of the optimal parameters. Overall, the results lie close together for all parameter sets, which means that there was not too much overfitting to a particular word during validation.

However, it is interesting to observe that the performance of handwriting evaluation is different for the three words. This motivates further studies to determine what kind of words, or characters, are best suited to automatically assess the learning progress of children with respect to legibility.

Table 4. Classification accuracy on the test set with three classes.

Parameters	Lundi	Lunes	Plumes
P1	0.753	0.693	0.667
P2	**0.822**	**0.730**	**0.687**
P3	0.810	0.720	0.657

Table 5. Classification accuracy on the test set with two classes.

Parameters	Lundi	Lunes	Plumes
P1	0.776	0.847	0.758
P2	**0.822**	**0.850**	**0.788**
P3	0.816	0.843	0.747

With respect to the first objective of our study (O1), we can summarize that the automatic evaluation of legibility corresponds well to human judgement but it leaves room for improvements regarding the classification accuracy.

4.3 Results with Clustering

To evaluate if the graph-based approach leads to meaningful groupings of the handwritten words, we focus on one of the words, “lundi”, and use the same set of optimized meta-parameters that was established for the task of classification.

The Silhouette index [7,24] to assess the quality of the clustering produced. It is therefore possible to compare the results of clustering methods and algorithms, and also to determine the number of clusters to retain [4,5]. Figure 5 shows the evolution of the silhouette index as a function of the number of clusters k and the value of the index is higher for $k = 2$.

The next step is to analyse the content of the clusters produced using the classes assigned by the experts. Table 6 shows the frequency of ground truth labels in the clusters with $k = 2$ for all clustering methods. In order to simplify understanding of the table, the clustering algorithms producing the same distribution statistics have been grouped together (from A1 to A4).

A1 produces two clusters, gathering 75% of the words marked legible in C2 and up to 75% of the not very legible and illegible in C2. Algorithms A2 to A4 do not produce significant results, yet their silhouette index values are higher overall than those of A1.

With respect to the second objective of our study (O2), we can summarize that the graph-based dissimilarity leads to a generally good clustering quality. However, the legibility alone cannot explain the groupings that result from graph-based matching.

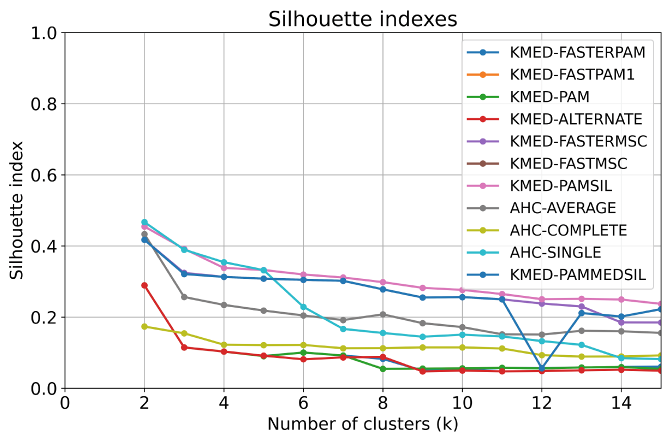

Fig. 5. Evolution of the silhouette index as a function of the number of clusters k from 2 to 15 for “lundi”; KMED and AHC algorithms were used with different configurations (PAM, FASTERPAM, AVERAGE, COMPLETE, etc.).

Table 6. Frequency of expert classes per cluster for each group of clustering algorithms for $k = 2$ (with A1: AHC-COMPLETE; A2: KMED-FATERPAM, KMED-FASTPAM1, KMED-PAM, KMED-ALTERNATE; A3: KMED-FASTERMSC, KMED-FASTMSC, KMED-PAMMEDSIL; A4: AHC-AVERAGE, AHC-SINGLE, KMED-PAMSIL).

Algorithm	Cluster	Expert classes (EC)		
		Legible	Not very legible	Illegible
A1	C1	0.255	**0.75**	**0.667**
	C2	**0.745**	0.25	0.333
A2	C1	0.34	0.333	0.333
	C2	0.659	0.667	0.667
A3	C1	0.021	0.083	0.048
	C2	0.979	0.917	0.952
A4	C1	0.007	0.0	0.0
	C2	0.993	1.0	1.0

5 Conclusion

In this paper, we have investigated graph-based representation of handwriting and graph matching for performing a visuo-structural evaluation of handwriting with respect to legibility.

The experimental evaluation demonstrates that the automatic method is well related to the judgment of human experts. For the two-class problem between legible and illegible, we report a classification accuracy between 79–85% depending on the word. For the three-class problem between legible, not very legible, and illegible, the performance drops but we still achieve an accuracy between 69–82%. Our clustering experiments have demonstrated that the graph-based similarity leads to clear groups of words but legibility on its own cannot explain these groupings.

It is noteworthy that the ground truth itself is ambiguous in the sense that also human experts tend to disagree on the legibility. In future work we aim to further improve the quality of the ground truth by including a larger number of experts. Such challenge can be supported by the crowdsourcing function of Copilotr@ce that has already been used by two of the three experts for this present study. Furthermore, we would like to include more diverse words in our study, as the accuracy varies among different words. It would also be interesting to perform the analysis at the level of patterns like characters rather than words, and to investigate what kind of words and characters are best suited to assess the handwriting quality. Finally, we would like to highlight that we have focused our investigation on one particular type of graph, keypoint graphs, and one particular type of graph matching, the Hausdorff edit distance. Thus, these non-language-dependent choices imply that our method is a priori suitable for any spelling and linguistic system and not only those of the French language. A promising line of research would be to investigate and compare other representation and matching paradigms in more detail for other languages such as, for example, Creole, English, Spanish which are those of the many allophone students schooled in Guadeloupe.

References

1. Bara, F., Gentaz, É., Colé, P.: Comment les enfants apprennent-ils à écrire et comment les y aider. Apprentissages et enseignement. Sciences cognitives et éducation, 9–24 (2006)
2. Bara, F., Morin, M.-F., Montésinos-Gelet, I., Lavoie, N.: Conceptions et pratiques en graphomotricité chez des enseignants de primaire en france et au québec. Revue française de pédagogie. Recherches en éducation (176), 41–56 (2011)
3. Barnett, L., Anna, M.P., Rosenblum, S.: Development of the handwriting legibility scale (HLS): a preliminary examination of reliability and validity. Res. Dev. Disabil. **72**, 240–247 (2018)
4. Biabiany, E., Bernard, D.C., Page, V., Paugam-Moisy, H.: Design of an expert distance metric for climate clustering: the case of rainfall in the lesser Antilles. Comput. Geosci. **145**, 104612 (2020)
5. Biabiany, E., Page, V., Bernard, D.C., Paugam-Moisy, H.: Using an expert deviation carrying the knowledge of climate data in usual clustering algorithms. In: CAP and RFAIP Joint Conferences, Vannes, May 2020
6. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty years of graph matching in pattern recognition. Int. J. Pattern Recogn. Artif. Intell. **18**(3), 265–298 (2004)
7. Amorim, R.C.D., Hennig, C.: Recovering the number of clusters in data sets with noise features using feature rescaling factors. Inf. Sci. **324**, 126–145 (2015)
8. Erez, N., Parush, S.: The Hebrew handwriting evaluation. School of Occupational Therapy. Faculty of Medicine. Hebrew University of Jerusalem, Israel (1999)
9. Fischer, A., Riesen, K., Bunke, H.: Graph similarity features for HMM-based handwriting recognition in historical documents. In: Proceedings International Conference on Frontiers in Handwriting Recognition, pp. 253–258 (2010)
10. Fischer, A., Suen, C.Y., Frinken, V., Riesen, K., Bunke, H.: Approximation of graph edit distance based on Hausdorff matching. Pattern Recogn. **48**(2), 331–343 (2015)
11. Florence, B., Nathalie, B.-B.: Handwriting isolated cursive letters in young children: effect of the visual trace deletion. Learn. Instr. **74**, 101439 (2021)
12. Fogel, Y., Rosenblum, S., Barnett, A.L.: Handwriting legibility across different writing tasks in school-aged children. Hong Kong J. Occup. Ther. **35**(1), 44–51 (2022)
13. Hamdi, Y., Akouaydi, H., Boubaker, H., Alimi, A.M.: Handwriting quality analysis using online-offline models. Multimedia Tools Appl. **81**(30), 43411–43439 (2022)
14. Hamstra-Bletz, L., DeBie, J., Den Brinker, B.P.L.M., et al.: Concise evaluation scale for children's handwriting. Lisse Swets **1**, 623–662 (1987)
15. Larsen, S.C., Hammill, D.D.: Test of legible handwriting (Pro-Ed, Austin, TX) (1989)
16. Lenssen, L., Schubert, E.: Clustering by direct optimization of the medoid silhouette. In: Skopal, T., Falchi, F., Lokoč, J., Sapino, M.L., Bartolini, I., Patella, M. (eds.) Similarity Search and Applications, pp. 190–204. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17849-8_15

17. Li, T., Rezaeipanah, A., El Din, E.M.T.: An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. J. King Saud Univ. Comput. Inf. Sci. **34**(6, Part B), 3828–3842 (2022)
18. Maergner, P., et al.: Combining graph edit distance and triplet networks for offline signature verification. Pattern Recogn. Lett. **125**, 527–533 (2019)
19. Phelps, J., Stempel, L.: Handwriting: evolution and evaluation. Ann. Dyslexia **37**, 228–239 (1987)
20. Rémi, C., Nagau, J.: Copilotrace: a platform to process graphomotor tasks for education and graphonomics research. In: Carmona-Duarte, C., Díaz, M., Ferrer, M.A., Morales, A. (eds.) Intertwining Graphonomics with Human Movements - 20th International Conference of the International Graphonomics Society, IGS 2021, Las Palmas de Gran Canaria, Spain, 7–9 June 2022, Proceedings. LNCS, vol. 13424, pp. 129–143. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19745-1_10
21. Riba, P., Lladãs, J., Fornés, A.: Handwritten word spotting by inexact matching of grapheme graphs. In: Proceedings 13th International Conference on Document Analysis and Recognition, pp. 781–785 (2015)
22. Rosenblum, S., Parush, S., Weiss, P.L.: Computerized temporal handwriting characteristics of proficient and non-proficient handwriters. Am. J. Occup. Ther. **57**(2), 129–138 (2003)
23. Rosenblum, S., Weiss, P.L., Parush, S.: Product and process evaluation of handwriting difficulties. Educ. Psychol. Rev. **15**, 41–81 (2003)
24. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. **20**, 53–65 (1987)
25. Schubert, E., Lenssen, L.: Fast K-medoids clustering in rust and Python. J. Open Source Softw. **7**(75), 4183 (2022)
26. Schubert, E., Rousseeuw, P.J.: Fast and eager K-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. Inf. Syst. **101**, 101804 (2021)
27. Scius-Bertrand, A., Studer, L., Fischer, A., Bui, M.: Annotation-free keyword spotting in historical Vietnamese manuscripts using graph matching. In: Proceedings International Workshop on Structural and Syntactic Pattern Recognition (SSPR) (2022)
28. Soppelsa, R., Albaret, J.-M.: Evaluation de l'écriture chez l'adolescent. le bhk ado. Entretiens de Psychomotricité, 66–76 (2012)
29. Stauffer, M., Fischer, A., Riesen, K.: Graph-Based Keyword Spotting. World Scientific (2019)
30. Vinter, A., Chartrel, E.: Effects of different types of learning on handwriting movements in young children. Learn. Instr. **20**(6), 476–486 (2010)