The Neglected Role of GUI in Performance Evaluation of AI-Based Transcription Tools for Handwritten Documents

Giuseppe De Gregorio(✉) and Angelo Marcelli

DIEM, University of Salerno, Via Giovanni Paolo II 132, 84084 Fisciano, SA, Italy
{gdegregorio,amarcelli}@unisa.it

Abstract. This paper aims to inspect the often neglected role of Graphical User Interfaces (GUI) in AI-based tools designed to assist in the transcription of handwritten documents. While the precision and recall of the handwritten word recognition have traditionally been the primary focus, we argue that the time parameter associated with the GUI, specifically in terms of validation and correction, plays an equally crucial role. By investigating the influence of GUI design on the validation and correction aspects of transcription we want to highlight how the time that the user must take to interact with the interface must be taken into account to evaluate the performance of the transcription process. Through comprehensive analysis and experimentation, we illustrate the profound impact that GUI design can have on the overall efficiency of transcription tools. We demonstrate how the time saved through the utilization of an assistant tool is heavily dependent on the operations performed within the interface and the diverse features it offers. By recognizing GUI design as an essential component of transcription tools, we can unlock their full potential and significantly improve their effectiveness.

Keywords: Handwritten · Document Transcription · Document Analysis · Historical Document Processing

1 Introduction

In an increasingly digital world, the task of converting handwritten documents into a digital format can be time-consuming and challenging. The rapid advancements in artificial intelligence (AI) have paved the way for innovative solutions in various fields, including transcription [9]. AI tools that focus on transcribing handwritten text offer immense potential for increased efficiency and accuracy, potentially revolutionizing how we manage handwritten documents.

The utilization of AI tools for transcription purposes involves leveraging sophisticated algorithms, neural networks, and machine learning techniques to interpret and convert handwritten text into digital form. These tools learn from

A. Parziale et al. (Eds.): IGS 2023, LNCS 14285, pp. 151–164, 2023.
https://doi.org/10.1007/978-3-031-45461-5_11

vast amounts of data, acquiring the ability to recognize patterns, characters, and words, enabling them to accurately transcribe handwritten documents with increasing precision.

This technology has the potential to significantly streamline workflows, improve accessibility, and facilitate data analysis. Indeed, one of the primary benefits of using AI tools for transcribing handwritten documents is the potential for significant time savings [5]. What used to take hours or even days to manually transcribe can now be accomplished in a fraction of the time. This increased efficiency not only enhances productivity for individuals and organizations but also allows for expedited access to critical information contained within handwritten documents.

However, it is essential to consider the limitations and challenges associated with using AI tools for transcription. Handwriting can vary significantly between individuals, making it difficult for AI systems to accurately interpret unique styles and idiosyncrasies. Complex or degraded handwriting, smudges, or unclear markings can further compound the challenge. Additionally, certain languages or scripts pose additional difficulties, as AI tools may be primarily trained on specific languages or character sets. An important example is given by handwritten documents of historical interest [7]. Working with handwritten historical documents poses unique challenges. The passage of time, exposure to the elements, and ageing of materials can cause deterioration, making the texts difficult to read or comprehend. The use of archaic language, abbreviations, and unique writing conventions prevalent in different time periods can also pose challenges for contemporary readers and researchers.

The use of AI tools for transcribing can facilitate the digitization and transcription process. These tools can assist in deciphering handwriting, enhancing legibility, and converting the content into searchable digital formats, making the documents more accessible to researchers and the general public [3,18]. However, the process of assisted transcription raises questions and considerations regarding the accuracy and the role of human involvement. While AI models have made significant progress, errors can still occur, especially when confronted with ambiguous or illegible handwriting. It is crucial to approach AI-transcribed documents with caution and consider essential the need for human intervention or verification to ensure accuracy and reliability.

Traditionally, the primary emphasis in transcription systems has been on achieving high accuracy rates in recognizing handwritten words. While this is undoubtedly important, it is equally essential to recognize the equally critical role played by the Graphical User Interfaces (GUI), particularly in terms of validation and correction processes. The time parameter associated with these GUI interactions can significantly impact the overall efficiency of transcription tools.

Given the need for user intervention to ensure an accurate and error-free document transcript, regardless of the AI technology employed, human-machine interaction plays a significant role. Consequently, the time saved by using this system does not simply depend on the performance of the AI model used and its ability to recognize handwritten text after proper training; it is equally important to consider how quickly users can interact with the system interface.

The user must verify that the AI tool has accurately associated transcriptions with the images of words present in the documents to be transcribed. All transcriptions correctly linked should be validated. Furthermore, any mistakes made by the text recognition system must be corrected by replacing wrong transcriptions with accurate ones. Lastly, if the AI system was unable to recognize any words, a transcript must be provided manually for them.

All of these operations take time, and the amount of time is contingent on the design choices of the interface, what basic operations were chosen to interact with it, and how many and which features are available to the user. Intuitively, it makes sense that these operations should require less time than manually transcribing a word since using the entire system can lead to faster total transcription. It is more difficult to grasp what impact each element has on obtaining an overall reduction in the time gain. In this work, we focus our attention on trying to determine how much time can be dedicated to validating and correcting output from text recognition systems while sustaining the decrease in transcription time.

The paper is then organized as follows: in Sect. 2, we provide a detailed overview of the transcription process when it is assisted by a Keyword Spotting (KWS) system, emphasizing the time course of the typical interaction between the user and the validation/correction interface; in Sect. 3, we present the experimental results obtained from three datasets containing handwritten documents from the 13th to 18th centuries; while, in Sect. 4, we discuss the experimental findings; lastly, in Sect. 5, we draw some preliminary conclusions and outline future investigations.

2 The Transcription Process

In the process of transcribing a set of handwritten documents, a Keyword Spotting system can be utilized to reduce the user workload. A KWS system is a machine learning tool that has the charge of locating words it knows how to represent within images of handwritten pages. The preparatory phase requires creating a training set, hereinafter referred to as TS, containing the representation of each word image (in terms of a suitable set of features) and its correct transcription. For the sake of performance, it is usual that a smaller portion of the total collection is used, and it is crucial that an accurate and complete transcription of TS is available; when it does not exist, it is up to the user to manually transcribe selected documents for use in TS. In such a case the user must spend the time t_{man} to read a word of the document and type-in the transcript. Thus, t_{man} depends mostly on the proficiency of the user in reading and providing the transcript.

For the transcription of the rest of the collection, hereinafter referred to as the data set and denoted by DS, the Keyword Spotting system can be utilized to retrieve words that are most similar in representation to those of the keyword list. As such, it is possible to recover transcripts for keywords within DS without the system having to explicitly recognize the text present in the images. This allows KWS systems to be robust when dealing with manuscript collections consisting of a small number of documents [1].

Ultimately, the goal of the system is to accurately transcribe the list of word images present in DS, so that manual transcription is no longer necessary, thus saving user time and effort. High values of the KWS system Precision p and Recall r would further optimize the transcription process, as a greater number of correctly identified words yield more savings in terms of time required for transcribing a collection. Consequently, it is important that KWS systems strive for an optimal performance output so as to achieve maximum efficiency in obtaining an accurate transcription.

The performance of a KWS system, is given in terms of its p and recall r, and since they are both smaller than 1, the KWS is liable for mistakes in spotting the word image corresponding to the keyword of the query, thus providing the wrong transcript, as well as for missing some words, thus being unable to transcribe all of the words of DS. Additionally, KWS systems can struggle with the problem of out-of-vocabulary (OOV) words, i.e. words present in DS but not included in TS, resulting in either no spotting at all or a significant drop in performance. Consequently, it is necessary to incorporate a validation stage to guarantee that all the words in the documents are accurately transcribed. This includes verifying the output of the KWS system, confirming correct transcriptions, correcting errors, and manually transcribing any missed word.

This user-system interaction necessitates a Graphical User Interface that enables the user to view the image of the word to transcribe in addition to the list of transcription options generated by the KWS system from which to choose the correct one, thus spending the time t_{val} to achieve the transcription of the word. Thus, t_{val} is contingent on how the GUI was constructed and which operation is dedicated to validate the correct transcript. As an example, one could envision validating the output by clicking on the right transcription with a mouse, or using the arrow keys on the keyboard to pick out the correct entry from the list, or interacting directly with the interface through a touchscreen device. Moreover, if the list is ranked according to the likelihood of a word to be the right transcription and the default option is that the top-ranking element is the correct interpretation, whenever this happens to be true the correct transcription can be obtained by clicking a mouse button, or by pressing/touching a dedicated key. The time t_{val} depends also on how many transcription options are available to the users via the GUI; a quick scrolling is possible when there are few elements in the list, but having few alternatives decreases the likelihood of the correct transcription to be in the list. Figure 1 illustrates an example of an interface for assisting the user during the validation of system output. In the figure, the interface proceeds line by line, showing the current line of text on top of the screen, with the word to transcribe highlighted in the text line and displayed at the centre of the interface while on its right side is the list of possible transcriptions proposed by the system from which the user must choose the correct one.

When the system is unable to provide the right option in the transcription options list, or when it cannot produce any (which may occur when the word is an OOV word), the correct transcription must be provided manually. The

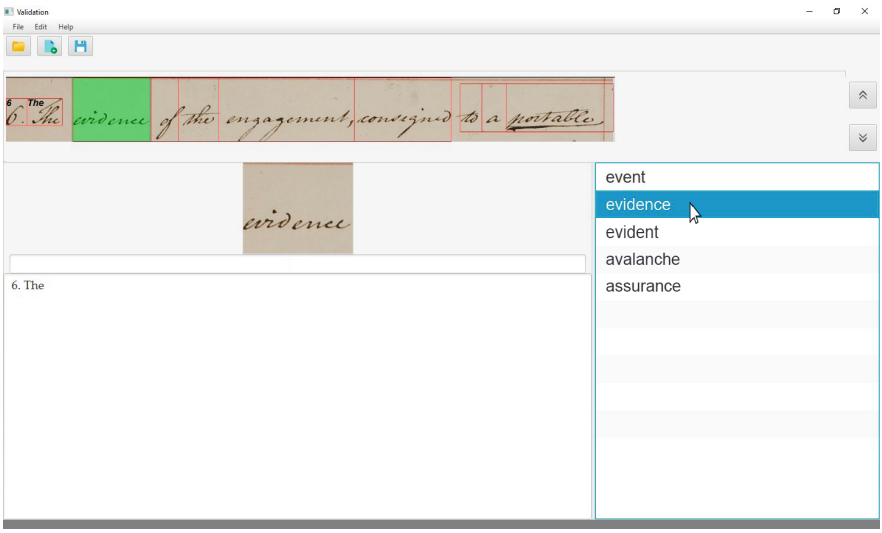

Fig. 1. An example of an interface during the validation process. The transcript of the current word "*evidence*" is present in the list of options and the user must simply validate it by selecting the correct transcript from the list.

time t_{cor} to perform this operation also depends on the features of the system interface. For instance, the interface can be fitted with an auto-complete mode that can expedite and accelerate the typing of the transcriptions. After typing in the initial letters of the correct transcription, the system can search for relevant keywords compatible with those same letters. At this point, the GUI could potentially offer up the correct transcription that can be selected without writing out all of it (Fig. 2). Similarly to the previous case, the use of the system is profitable when $t_{cor} \leq t_{man}$.

The parameters that define the interaction with the interface, thus, are t_{val} and t_{cor}. In short, the former represents the time taken for a user to view a handwritten word and determine if the correct transcription is among those presented by the system. The latter indicates how much time is required to enter a transcription manually when it has been determined that no valid alternative was supplied. Paying attention to these parameters when designing a validation/correction graphical interface can be of utmost importance. An interface that requires too much time for interactions may effectively cancel out the gain in time that the use of a KWS system provides. Additionally, interaction operations necessary for interface effectiveness, even when designed well, are not instantaneous. It is therefore important to consider how the time gain is influenced by the elementary operations required by a particular interface given the performance of the KWS system used. For these reasons, it is important to estimate the time gain G obtainable by the system depending on *both* the performance indexes of the KWS system r and p and on the time parameters of the interface t_{val} and t_{cor}:

$$G(p, r, t_{val}, t_{cor}) \tag{1}$$

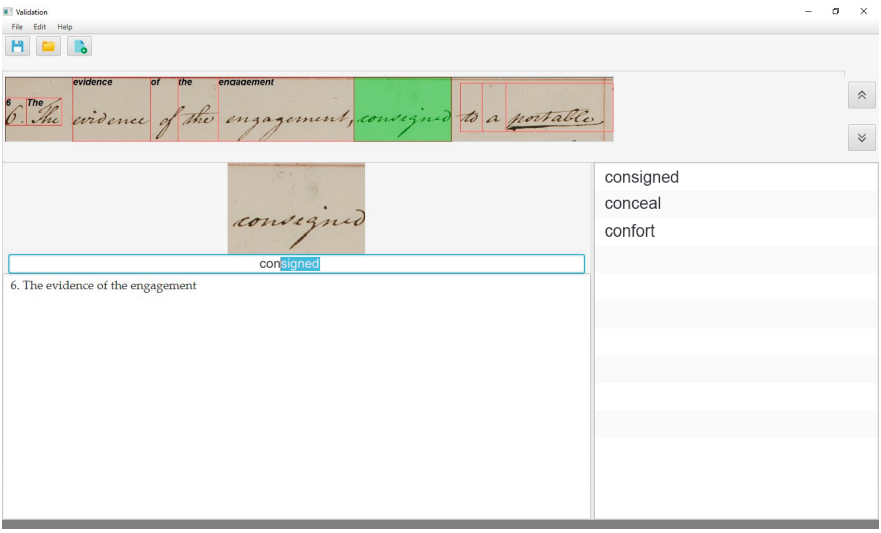

Fig. 2. The word "*consigned*" was not recognized by the system, so the user is forced to enter the transcript manually. The autocomplete system allows the identification of the correct transcription after the user has entered just three letters.

It is worth noting that when both t_{val} and t_{cor} are equal to t_{man} there is no gain, and that it increases as t_{val} and t_{cor} becomes smaller and smaller with respect to t_{man}.

3 Experimental Results

In this section, we experimentally assess the impact of the GUI temporal parameters t_{val} and t_{cor} on the temporal gains obtainable using a KWS to assist the transcription of an entire collection of handwritten documents. We first assess the performance of KWS in terms of the size of the training set TS. Once the size of TS has been established and the performance indices p and r determined, we then investigate how varying the time parameters t_{val} and t_{cor} impacts the resulting time gain G.

3.1 Datasets

For experimentation, three collections of historical handwritten documents commonly used as benchmarks for KWS systems [2,11,13,14,16,17] were taken into consideration, namely the George Washington dataset, [10], the Bentham dataset [12], and the Parzival dataset [6]. For the purpose of the experiments, only 20 pages extracted from the Bentham dataset were used, while the entire dataset was used in the remaining cases. The George Washington and the Bentham datasets both originate from the 18th century and are written in English by a

single writer, while the pages of the Parzival dataset are written in Middle High German and were produced by three authors in the 13th century.

Each dataset has been divided into the training set TS, composed of some pages of the collection, and the DS set made up of the remaining pages to be transcribed. Table 1 shows the details of the different datasets highlighting the number of pages and the number of words contained in each of them.

Table 1. Composition of datasets in terms of number of pages and number of words.

Dataset	Num Pages	Num Words
Washington	20	4819
Bentham	20	3478
Parzival	47	23412

3.2 KWS System

The Keyword Spotting System (KWS) we used is based on PHOCNet [15] and it was set up for segmentation-based Query-by-Example search (QbS). The images and transcriptions of the terms in the training set TS were used to train the PHOCnet. During query time, all distinct transcriptions from TS were taken and their corresponding PHOC representation was used as the keyword list. Bray-Curtis dissimilarity [4] was utilized to measure the similarity between images in the keyword list and the images of the words to be transcribed in the set DS. The KWS is able to return an ordered list of possible transcriptions for a query word image, and the order of the entries in the list is defined by the distance measured between the query word and the keywords.

3.3 KWS Performance

In the first experimental phase, the KWS system's performance in terms of Precision and Recall were assessed for each DS reported in Table 1. Figure 3 illustrates how varying the number of pages in TS affects Precision and Recall for each dataset. The experiments were executed three times for each dataset, randomly selecting the order of pages in TS each time, and the results are reported in terms of the average values.

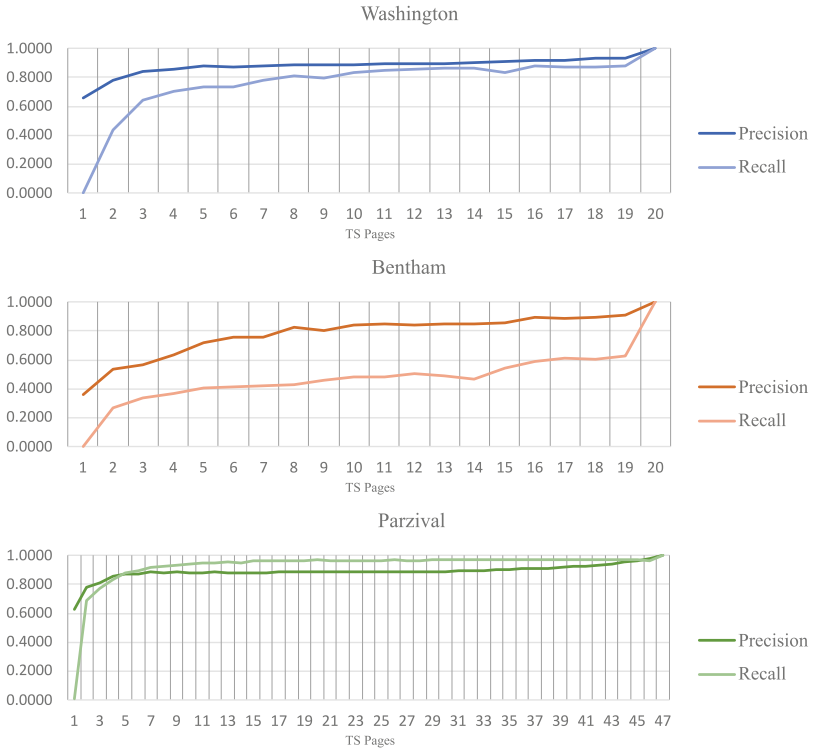

Fig. 3. Precision and Recall as the number of pages of the TS set varies on the three datasets.

3.4 How GUI Times Affects the Time Gain

As illustrated by the graphs in Fig. 3, the precision and recall value is contingent upon the number of pages in the training set TS and tends to remain steady for TS comprised of 5 to 10 pages. To analyze how the temporal gain of the transcription varies based on temporal-dependent user interface parameters, we use the values obtained from $TS = 5$ pages and $TS = 10$ pages.

The time parameters of the interface t_{val} and t_{cor}, along with the performance indices p and r of the KWS system, affect the time required to achieve a complete and accurate transcription of a collection of handwritten documents. Marcelli et al [8] derive this dependency by proposing a model that can estimate the time gain achievable utilizing a KWS system to support the transcription process; it is notable that the parameters for a lexicon-based KWS are similar to those considered in this study. We then employ the model introduced in [8] to compute the time gain and assess how different timescale values of the t_{val} and t_{cor} influence performance in terms of time gain. It is essential to note that the model predicts using precision and recall indices p_i and r_i computed for each keyword in the list. In this work, we assumed the same average precision and

recall values p and r for each keyword, as is commonly done in the literature. Similarly, we use for t_{man} the average instead of a different t_{man_i} for each word, and lastly, following the observation reported at the end of Sect. 2, express t_{val} and t_{cor} with respect to t_{man} rather than independently. By doing so, it will be possible to estimate the gain given the *implementation* of the interface by using their values measured in the preliminary phase, but also to set the time constraints for the *design* of the interface to achieve the desired estimated gain. We then adjust the two GUI time parameters t_{val} and t_{cor} from 0.1% to 100% of t_{man} and estimate the gain.

Figure 4, Fig. 5, and Fig. 6 illustrate the results of the Washington, Bentham, and Parzival datasets. The left panel of the figures displays the case in which TS is made of 5 pages, while the right panel shows the results obtained when TS is made of 10 pages. The graphs at the top of the figure illustrate how the values of t_{cor} and t_{val} vary when the temporal gain of the transcript is set to zero. The Zero Gain Line delineates which time parameters reset the gain, dividing the plane into two semi-planes. The area below the line is the positive gain area, that is the area in which a positive gain and therefore a reduction in transcription time is obtained, while above the line there is the negative gain area, which represents the area in which the transcription time is greater than the time necessary for a completely manual transcription. Moving downward further away from the Zero Gain Line the absolute value of temporal gain increases. This behaviour can be observed more clearly in the lower part of the figures, which highlights bands that link to gain range, which we will refer to as Time Gain Bands; only below the Zero Gain Line are bands with positive gains, and travelling further down from it increases the value of the temporal gain.

Observing the figures, it can be noted that with a TS of 10 pages, the area of positive gain increases, while the number of gain bands decreases. This implies that larger time gains are achievable when the TS consists of 5 pages, but stricter constraints for the interplay between validation and correction times must be enforced since the area for positive gain is reduced.

4 Discussion

Upon analyzing Figs. 4, 5, and 6, it becomes evident that regardless of the case, the positive gain area is more significant when using a training set TS composed of 10 pages compared to 5 pages. This observation suggests that achieving a positive gain becomes easier when working with a larger training set. This behaviour aligns with the trends exhibited by the performance indices of the Keyword Spotting system, as illustrated in Fig. 3. The larger the training set, the higher the potential for improving the KWS system's performance. Additionally, with a larger training set, the keyword list expands, while the number of pages requiring transcription decreases.

However, it is intriguing to note that when examining the Time Gain Band graphs, it is apparent that higher time gains can be achieved with a training set consisting of only 5 pages. Furthermore, for the case of $TS = 5$, once a

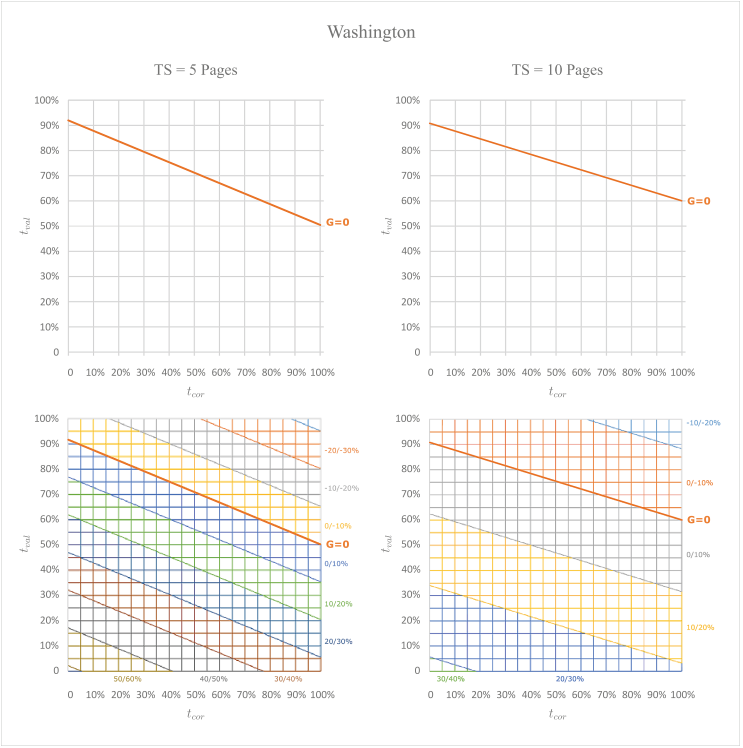

Fig. 4. The graphs above illustrate the Zero Gain Line in the t_{cor}/t_{val} plane, whereas the graphs below depict the varied Time Gain Bands for the Washington dataset. The axes values are expressed in percentage with respect to the manual transcription time of a word t_{man}.

desired time gain has been established, it is observed that a wider range of time parameters can lead to achieving this time gain. To clarify this behaviour, Table 2 presents the constraints that the validation time t_{val} must meet to achieve at least 10% and 20% of the time gain when the correction time t_{cor} is set at half the duration of manual transcription t_{man}. Notably, when the training set size TS is smaller, there is more flexibility in terms of the validation time allowed for achieving the target gain.

These findings indicate that while a larger training set generally leads to more favourable outcomes in improving the KWS system performance and obtaining a larger positive gain area, utilizing a smaller training set can result in higher time gains. The Time Gain Band analysis highlights the range of validation times that can be considered while achieving a desired time gain, especially when working with a smaller training set. This information underscores the importance of carefully selecting the appropriate training set size and considering the associated validation and correction times to optimize time gains in the transcription process.

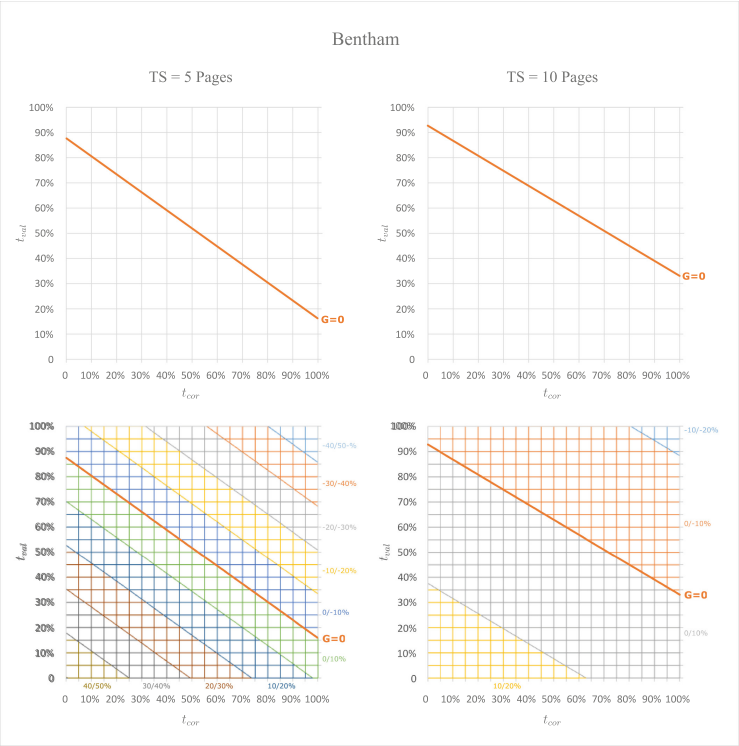

Fig. 5. The graphs above illustrate the Zero Gain Line in the t_{cor}/t_{val} plane, whereas the graphs below depict the varied Time Gain Bands for the Bentham dataset. The axes values are expressed in percentage with respect to the manual transcription time of a word t_{man}.

Table 2. Constrain on the validation time t_{val} when the t_{cor} is set at the half of ms and a time gain of at least 10% and at least 20% is desired.

Gain	Dataset	t_{val} ($\%t_{man}$)	
		$TS = 5$	$TS = 10$
10%	*Wasingthon*	<65%	<50%
	Benham	<35%	<10%
	Parzival	<60%	<50%
20%	*Wasinghton*	<55%	<20%
	Bentham	<15%	N/A
	Parzival	<50%	<25%

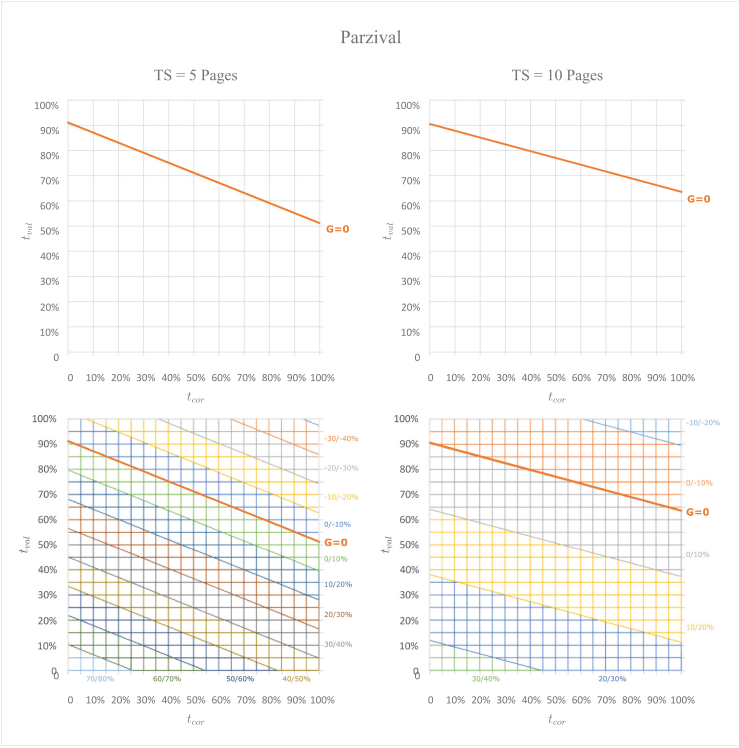

Fig. 6. The graphs above illustrate the Zero Gain Line in the t_{cor}/t_{val} plane, whereas the graphs below depict the varied Time Gain Bands for the Parzival dataset. The axes values are expressed in percentage with respect to the manual transcription time of a word t_{man}.

5 Conclusions

In conclusion, the study conducted sheds light on the impact of temporal parameters within the user interface of an assisted transcription system for handwritten documents. The findings demonstrate the crucial role these parameters play in determining the time saved through the utilization of such a system. By conducting various experiments and analyzing the results, the study establishes a clear link between the temporal parameters of the interface and the achievable time gain. It becomes evident that not only does the performance of the AI-based supporting machine learning tool contribute to reducing transcription time, but the design and functionality of the user interface also significantly influence the overall efficiency.

Moreover, the study implies that in situations where handwriting recognition systems fail to meet desired performance levels, it becomes essential to implement strategies aimed at minimizing interface interaction time in order to maintain a positive time gain. One possible strategy could involve limiting the

number of options presented during the validation phase, thereby expediting the process. This observation suggests a potential correlation between the performance indices of the keyword spotting system and the time parameters within the interface, especially when aiming to achieve a desired time gain. To gain further insights, future investigations should focus on clarifying this relationship, ultimately aiming to provide valuable observations and recommendations for the design and development of graphical interfaces used in assistance systems for handwritten document transcription.

By conducting more research in this domain, it will be possible to refine the design and functionality of the validation and correction interfaces. These improvements can lead to enhanced usability and efficiency, ultimately benefiting users of transcription assistance systems for handwritten documents. The study's findings offer valuable insights into the intricate interplay in human-computer interaction, hopefully paving the way for future advancements in this field.

References

1. Ahmed, R., Al-Khatib, W.G., Mahmoud, S.: A survey on handwritten documents word spotting. Int. J. Multimed. Inf. Retr. **6**, 31–47 (2017)
2. Aldavert, D., Rusiñol, M., Toledo, R., Lladós, J.: A study of bag-of-visual-words representations for handwritten keyword spotting. Int. J. Doc. Anal. Recognit. (IJDAR) **18**, 223–234 (2015)
3. Aqab, S., Tariq, M.U.: Handwriting recognition using artificial intelligence neural network and image processing. Int. J. Adv. Comput. Sci. Appl. **11**(7) (2020)
4. Bray, J.R., Curtis, J.: An ordination of the upland forest communities of southern wisconsin. Ecol. Monogr. **27**, 325–349 (1957)
5. Cordell, R.: Machine learning and libraries: a report on the state of the field. Library of Congress (2020)
6. Fischer, A., et al.: Automatic transcription of handwritten medieval documents. In: 2009 15th International Conference on Virtual Systems and Multimedia, pp. 137–142. IEEE (2009)
7. Lombardi, F., Marinai, S.: Deep learning for historical document analysis and recognition-a survey. J. Imaging **6**(10), 110 (2020)
8. Marcelli, A., De Gregorio, G., Santoro, A.: A model for evaluating the performance of a multiple keywords spotting system for the transcription of historical handwritten documents. J. Imaging **6**(11), 117 (2020)
9. Memon, J., Sami, M., Khan, R.A., Uddin, M.: Handwritten optical character recognition (OCR): a comprehensive systematic literature review (SLR). IEEE Access **8**, 142642–142668 (2020)
10. Rath, T.M., Manmatha, R.: Word spotting for historical documents. Int. J. Doc. Anal. Recogn. **9**(2–4), 139 (2007)
11. Retsinas, G., Louloudis, G., Stamatopoulos, N., Gatos, B.: Keyword spotting in handwritten documents using projections of oriented gradients. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 411–416. IEEE (2016)
12. Sánchez, J.A., Romero, V., Toselli, A.H., Vidal, E.: ICFHR 2014 competition on handwritten text recognition on transcriptorium datasets (HTRTS). In: 2014 14th International Conference on Frontiers in Handwriting Recognition, pp. 785–790. IEEE (2014)

13. Serdouk, Y., Eglin, V., Bres, S., Pardoen, M.: Keyword spotting using siamese triplet deep neural networks. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1157–1162. IEEE (2019)
14. Sfikas, G., Gatos, B., Nikou, C.: Semicca: a new semi-supervised probabilistic CCA model for keyword spotting. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1107–1111. IEEE (2017)
15. Sudholt, S., Fink, G.A.: Phocnet: a deep convolutional neural network for word spotting in handwritten documents. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 277–282. IEEE (2016)
16. Toselli, A.H., Vidal, E., Romero, V., Frinken, V.: Hmm word graph based keyword spotting in handwritten document images. Inf. Sci. **370**, 497–518 (2016)
17. Wicht, B., Fischer, A., Hennebert, J.: Deep learning features for handwritten keyword spotting. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 3434–3439. IEEE (2016)
18. Yadav, P., Yadav, N.: Handwriting recognition system-a review. Int. J. Comput. Appl. **114**(19), 36–40 (2015)