

Evaluating an Interactive-Predictive Paradigm on Handwriting Transcription: A Case Study and Lessons Learned

Luis A. Leiva, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal

ITI – Institut Tecnològic d'Informàtica
Universitat Politècnica de València
Camí de Vera s/n, 46022 - Valencia, Spain
{luileito,vromero,ahector,evidal}@iti.upv.es

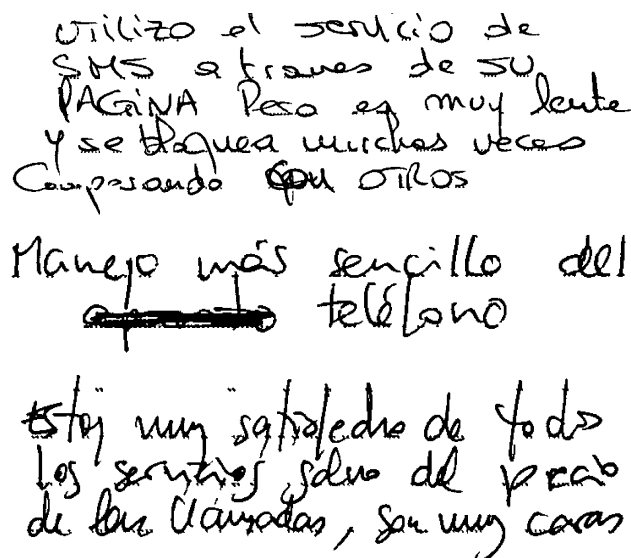
Abstract—Transcribing handwritten text is a laborious task which currently is carried out manually. As the accuracy of automatic handwritten text recognizers improves, post-editing the output of these recognizers could be foreseen as a possible alternative. Alas, the state-of-the-art technology is not suitable to perform this kind of work, since current approaches are not accurate enough and the process is usually both inefficient and uncomfortable for the user. As alternative, an interactive-predictive paradigm has gained recently an increasing popularity, mainly due to promising empirical results that estimate considerable reductions of user effort. In order to assess whether these empirical results can lead indeed to actual benefits, we developed a working prototype and conducted a field study remotely. Thirteen regular computer users tested two different transcription engines through the above-mentioned prototype. We observed that the interactive-predictive version allowed to transcribe better (less errors and fewer iterations to achieve a high-quality output) in comparison to the manual engine. Additionally, participants ranked higher such an interactive-predictive system in a usability questionnaire. We describe the evaluation methodology and discuss our preliminary results. While acknowledging the known limitations of our experimentation, we conclude that the interactive-predictive paradigm is an efficient approach for transcribing handwritten text.

Keywords—Handwriting Recognition; Interactive Transcription; field study;

I. INTRODUCTION

Since the introduction of computer machinery, the idea of fully automatic systems that would completely substitute the humans in certain types of tasks has gained an increasing popularity. In particular, scientific and technical research in the pattern recognition (PR) area traditionally have followed the *full automation* paradigm, even though, in practice, that approach often proves elusive or unnatural in many applications where technology is expected to assist rather than replace the human agents. Thus, the design paradigm of PR systems has recently experimented a notable shift towards systems where the decision process is affected by human feedback. One remarkable PR example where this feedback is successfully employed is the transcription of handwritten documents, which is currently becoming an important research topic.

Nowadays, there is an increasing number of applications



UTILIZO el servicio de
SMS a traves de SU
PAGINA Pero es muy lento
y se bloquea muchas veces
Comparando con otros
Manejo más sencillo del
~~aplicativo~~ teléfono
Estoy muy satisfecho de todos
los servicios salvo del precio
de las llamadas, son muy caras

Figure 1. An example of spontaneous, unconstrained handwritten text (in this case, a handwritten telephone survey response in Spanish), with a strongly irregular calligraphy, smears and cross-out words. Features like these make the handwriting recognition a non-trivial task.

which entail the transcription of handwritten documents into a textual electronic format for facilitating their posterior processing and efficient storage. This is for example the case of digital libraries that are publishing large quantities of digitized documents, which remain waiting to be transcribed [1]; or the case of spontaneous handwritten survey forms which require a prior recognition of their contents to extract then the relevant information (Figure 1.)

Actually, the problem of transcribing handwritten images can become a very laborious and expensive work, specially when the writing style of the text to transcribe becomes too variable. For instance, in the case of historic documents transcription this work is usually carried out by paleography experts who are specialized in reading ancient scripts; characterized, among other things, by different calligraphy and print styles from diverse places and time periods.

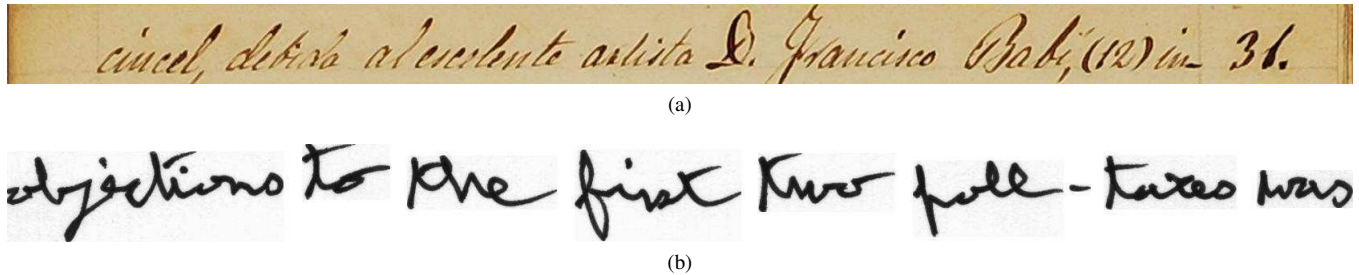


Figure 2. Sample lines extracted from the Cristo Salvador corpus, an ancient Spanish book of the XIX century (2a, see also Figure 3) and the IAM database (2b), available at <http://www.iam.unibe.ch/fki/databases/iam-handwriting-database>.

A. Background

State-of-the-art *cursive* handwritten text recognition (HTR) systems have been proved useful for restricted applications that involve form-constrained handwriting (such as bank check legal amounts) and/or fairly limited vocabulary (e.g., postal addresses). However, in the case of unconstrained handwritten documents (such as the before-mentioned cases of old manuscripts and spontaneous handwritten texts), current HTR technology typically only achieves results which are far from being directly acceptable in practice. Consequently, once a fully automated recognition process of one document has finished, heavy human revision is required to really produce a high-quality transcription. Hence, the human transcriber is therefore responsible for verifying and correcting the mistakes made by the system. Given the high error rates involved, a postediting approach is both inefficient and uncomfortable for the human corrector.

Researchers and practitioners have proposed crowdsourcing to speed up the transcription process, in which crowds of people are enticed to perform work over the Internet. An interesting application of crowdsourcing to help transcribing text is reCAPTCHA [2]. Unfortunately, this approach is an unreliable way to get the job done, i.e., nobody can assure that we will automatically get what we need, when we need it. Therefore, an interactive-predictive handwriting transcription (IHT) scenario was introduced to allow a more effective approach (see [1]).

Note that, in this work, handwriting data are given as images of scanned text; that is, a *static* information, generally referred to as *off-line* handwritten text. This is different from the *on-line* form of handwritten text, where handwriting data are given as a *temporal* sequences of coordinates — which generally allows online HTR systems to achieve much higher accuracy than offline HTR systems.

B. Interactive Handwriting Transcription

In IHT both the automatic HTR system and the human transcriber cooperate together to generate the final transcription of the text images. Under the above mentioned interactive-predictive paradigm, the user validates portions of a sentence (named *prefixes*) which are then used by the

IHT system to *predict* suitable continuations of the input sentence (named *suffixes*). The rationale behind this approach is to combine the accuracy provided by the transcription expert with the efficiency of the HTR system.

Human feedback signals in interactive systems rarely belong to the same domain as the one the main data stream comes from, thereby entailing some sort of multimodality. Of course, this is actually the case of IHT, where the main data are text images and feedback consists of keystrokes and/or pointer positioning actions.

The IHT paradigm follows ideas that have been already studied in the fields of speech recognition and machine translation [3]. Due to this paper's nature, instead of giving an extensive review on the IHT technology the reader is redirected to previous work such as [1], [4], [5]. Here we will focus on the evaluation of such an IHT paradigm, which was never done before with real users in a real setting — the related literature uses test-set-based *estimates* of user effort reduction, and we believe that this should be pragmatically verified.

II. EVALUATION METHODOLOGY

In order to test the effectiveness of the IHT technology several experiments were carried out on different cursive handwritten tasks such as old manuscripts [6], [7], handwritten answers from survey forms in modern Spanish [8] and handwritten full sentences in modern English [1], [9]. In all cases, empirical tests on these corpora based on annotated test-set data suggested that, using the IHT approach, considerable amounts of user effort could be saved with respect to pure manual work (non-interactive processing). While, of course, no definitive conclusions could be derived from these empirical tests, they clearly raised great expectations about the effectiveness and usability of this kind of interactive HTR technology. Therefore, in order to assess whether such expectations were in the right direction, we conducted a preliminary field study that compared the theoretical results with real users working with different implemented transcription engines.

A. Assessment Measures

In interactive PR systems the importance of the traditional recognition error rates is diminished, since the intention is to measure how well the user and the system work together. For this reason, to better judge the quality of the user transcriptions we used two objective test-set-based measures: word error rate (WER)¹ and word stroke ratio (WSR)². Both metrics have proven to be useful for estimating the reduction in human effort that can be expected by using IHT with respect to using a conventional HTR system [1]. On the other hand, in our subjective tests with real users we measured the time needed to fully transcribe each page with the different transcription possibilities (manual and IHT) as well as the residual WER (rWER)³ after human transcription — this value is expected to be greater than zero due to user’s errors.

B. Corpus

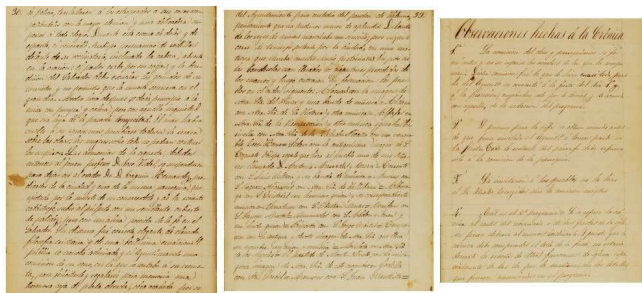


Figure 3. Examples of pages from the tested corpus. It is an old Spanish manuscript which shares many features with other ancient records, such as inconsistent spelling or an ornamented calligraphy.

The test corpus that we have used on the real user experiments was compiled from a XIX century handwritten document identified as “Cristo Salvador” (CS), which was kindly provided by the Biblioteca Valenciana Digital (Figure 3) [6]. This corpus is a legacy document which suffers the typical degradation problems of this kind of archives [10]. Among these, one can find the presence of smear, significant background of big variations and uneven illumination, spots due to humidity, and marks resulting from the ink that goes through the paper (generally called bleed-through). In addition, other kind of difficulties appear in these pages, such as different character sizes, decoration symbols, underlines, etc. The combination of these problems make the transcription of this kind of documents a difficult process for the system *and* also for the user.

¹The WER counts the minimum number of word-editing operations between the transcription proposed by the system and the reference transcription.

²The WSR is defined as the number of (word level) user interactions that are necessary to achieve the reference transcription, divided by the number of reference words.

³The rWER is the WER which remains after the user has typed the transcriptions or corrected/accepted the transcriptions proposed by the system.

It is important to remark that this corpus has a quite small training ratio (around 2.8 training running words per lexicon-entry). This results in undertrained language models, which will clearly increase the difficulty of the recognition task for the system.

C. Participants

It is worth mentioning that the cost of a formal field study of this kind of tasks is exceedingly high, since it typically involves expensive work by a panel of experts (usually paleographers). For that reason, we decided to start doing a preliminary exploration, recruiting regular computer users instead. Fourteen participants from our Computer Science department volunteered to cooperate, aged 28 to 61 (M=37.3, 3 females). Most of them were knowledgeable with handwriting transcription tasks, although none was a transcriber expert. Additionally, only three participants were aware of the existence of a transcription prototype prior to the evaluation. One user could not finish the evaluation, so the end user sample was 13 subjects.

D. Apparatus

Currently there are no commercial systems to assist the user in unconstrained handwriting transcription. Also, the few available prototypes found in the research literature are not publicly available. Luckily, two years ago we developed a web-based demo [11] to showcase the ITH technology. Therefore, we modified that system to carry out the field study. We implemented two HTR engines to assist the document transcription: a trivial manual system and our IHT system. The user interface (UI) was common to both engines. It provided basic text-editing capabilities, and allowed to transcribe interactively each page line by line (Figure 4). The application supported some special editing operations (such as insertions, rejections, or deletions) in order to better assist the user in the transcription process.

Also, a logging mechanism was embedded in the web application. It allowed us to register all user interactions in a fine-grained level of detail (e.g., keyboard and mouse events, client/server messages exchanging, etc.). The generated log files were reported in XML format for later postprocessing.

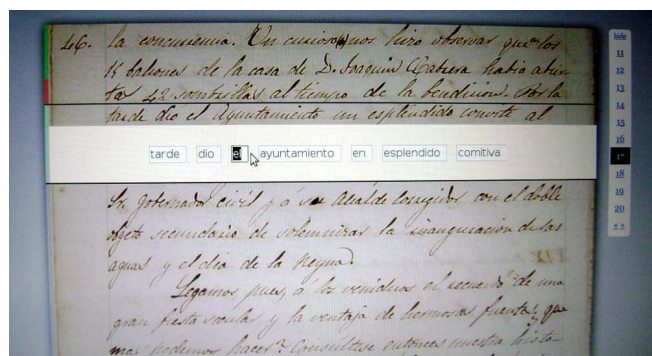
E. Procedure

Initially, the digitized images (i.e., the corpus pages) were automatically preprocessed and divided into lines. The results were visually inspected and the few line-separation errors (around 4%) were manually corrected. Table I summarizes the partition into training and test sets, which was used for the field study. The reference transcriptions were also available, containing 10918 running words with a vocabulary of 3287 different words.

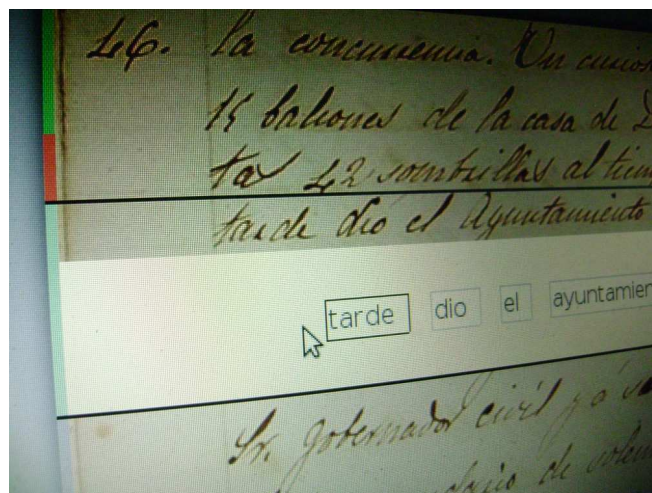
Participants accessed the web-based application via a special URLs that were sent to them by email. In order to familiarize with the UI, users informally tested each



(a)



(b)



(c)

Figure 4. IHT user interface. The book to transcribe was divided in pages (4a), and on each page the user could select one page at a time. Once a user clicks on a text segment, the main application unfolds the keyboard input area (4b). In the left margin of each page (4c), lines are marked as validated (all the text was reviewed), pending (only a fraction of the current text segment has been transcribed), or locked (someone else is working on the same text segment), respectively. For more details about the UI or the application workflow, the reader is redirected to [11].

Table I
DATA FROM THE HARD PARTITION OF CS CORPUS.

Set	Pages	Lines	Words	WER	WSR
training	33	675	6277	–	–
test	20	497	4691	33.5	32.1
user-test	2	32	477	24.5	23.0

transcription engine with some test pages, different from the ones reserved for the user-test. Then, people transcribed the two user-test pages; each one with both transcription engines. The two user-test pages selected for the field study had very similar test-set-based performance metrics (page 45: WER = 24.35%, WSR = 23.07%; page 46: WER = 24.69%, WSR = 23.04%; respectively), being also around the median value of the test-set.

It is important to remark that nobody saw both pages before the study. For that reason, it is clear that the engine that were tested at first would lead to poorer results — in the next trial users would need less effort in reading the image lines. Thus, to avoid possible biases due to human learnability, the first page (#45) was initially transcribed with the manual engine first; then the order was inverted for the second page (#46). Finally, participants filled out an online System Usability Scale (SUS) questionnaire [12] for both systems. Such online form included a text field to allow the users to submit free comments and ideas about their testing experience, as well as giving some insights about possible enhancements and/or related applicability.

F. Design

We carried out a within-subjects repeated measures design. We tested two conditions: transcribing a page with the manual and the IHT system, taking into account that each one was tested twice — to compensate the above-mentioned learnability bias. We performed a non-parametric test in each case, since normality assumptions did not hold (see below). Additionally, we studied if there were any correlation between trials and between measures variables, for such gathered performance metrics (time, residual WER, and WSR, respectively). We used the R Language [13] to process the data.

III. RESULTS AND DISCUSSION

In sum, we can assert that regarding *effectiveness* there are no significant differences, as expected, i.e., users can achieve their goals with any of the tested systems. However, in terms of *efficiency* the IHT system is the better choice. Regarding to *user satisfaction*, IHT again seems to be the most preferable option. Now let us delve into a more detailed analysis in order to shed more light to the obtained results. Initially we report the amount and nature of the differences found between both groups. Then we study both the statistical significance and the correlation between the measured variables.

A. Quantitative Analysis

Table II reports the result of the field study. We must emphasize that the daily use of any system designed to assist handwriting transcription would involve not having seen previously any of the pages (users usually read a page once and at the same time they just transcribe it).

Table II
MEAN (AND SD) PER PAGE FOR THE MEASURED VARIABLES: TIME (IN MINUTES), rWER & WSR (IN %), AND DIFFERENCES (IN %).

System		Time	rWER	WSR
Overall	Manual	11.1 (3.5)	8.6 (8.2)	97.8 (6.0)
	IHT	10.3 (3.7)	6.5 (3.7)	30.4 (6.1)
Difference		7.2	24.4	68.9
Page 45	Manual	12.8 (3.5)	12.8 (9.5)	97.3 (7.0)
	IHT	8.6 (3.2)	7.0 (4.1)	28.6 (4.1)
Difference		32.8	45.3	70.6
Page 46	Manual	9.4 (2.9)	4.1 (2.0)	98.4 (4.6)
	IHT	12.0 (3.4)	6.0 (3.3)	32.1 (7.1)
Difference		21.6	31.6	67.3

We computed the difference between both systems as $\text{diff} = \left| \frac{m-i}{\max(m,i)} \right|$, being m and i each measured variable in the manual and interactive versions, respectively.

To determine if data could be assumed to be normally distributed, we run a Shapiro-Wilk normality test [14]. Given that in most cases (see Figure 6) the results of the normality test were statistically significant, the data could not be considered normal. We decided thus to use the (non-parametric) two-sample Kolmogorov-Smirnov test for the evaluation study. Additionally, we measured the *probability of improvement* (POI), which estimates if a system is *a priori* better than another for a given user [15].

1) *Analysis of Time*: We observed that, overall, there are no differences in transcription times ($D = 0.16$, $p = 0.75$, n.s.). In general, the system used in second place always achieved the best time, because the user already knew the text. The remarkable result is that when the user reads a page in first place the chosen engine is not determinant, because one must spend time to accustom to the writing style, interpreting the calligraphy, etc. In this case the POI of IHT with respect to the manual engine is 53%.

2) *Analysis of rWER*: Overall, IHT was the best choice regarding to residual WER ($D = 0.11$, $p = 0.99$, n.s.). Although the differences are not statistically significant, the interesting observation is that IHT is the most stable of the systems — even better than when using the manual engine on an already read page (see Table IV). We must recall that the more stable in rWER a system is, the fewer residual errors are expected and therefore a high quality transcription is guaranteed. In this case, considering the first time that the user reads a page, the POI of the IHT engine over the manual engine is 69%.

3) *Analysis of WSR*: Interestingly, the WSR when using the manual engine was below 100%, since there are inherent errors (some users were unable to correctly read all the lines). That means that some users wrote less words in their final transcriptions than they really should have written when using the manual engine. In both conditions IHT was the best performer, and differences were statistically significant ($D = 1$, $p < 0.001$). The POI of the IHT engine regarding the manual engine is 100%. This means that the number of words a user must write and/or correct under the IHT paradigm is always much lower than with a manual system. Additionally, this fact increases the probability of achieving a high-quality final transcription, since users perform fewer interactions and are prone thus to less errors. It is also interesting to note that, on average, the real WSR achieved by the participants is fairly close to the objective user-test based estimates for the same pages and even closer to the overall test-set estimates reported in Table I.

B. Qualitative Analysis

Regarding user subjectivity, the SUS scores could be considered normally distributed. Thus, a Welch two-sample t-test was employed to measure the differences between both groups. We observed a clear tendency in favor to IHT ($t(22) = 0.25$, $p = 0.80$, n.s.), since users generally appreciate the guidance of the IHT system to suggest partial predictions, considering the difficulty of the task proposed in the field study.

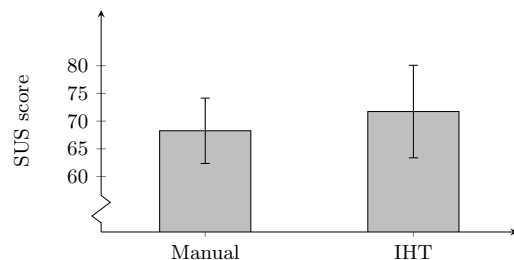


Figure 5. User satisfaction, according to the SUS questionnaire.

Most of the users' comments were alas related to the web UI rather than the transcription engines themselves. Some included "when clicking on a text field, the whole word is selected", "it is hard to remember some [keyboard] shortcuts", or "a clear and visual user manual would allow not having to learn almost anything before using the system." Additionally, four users complained about the segmentation of lines, which "made especially difficult reading those images where words had many ascenders/descenders." On the other hand, three users noticed that punctuation chars did not contribute to improve predictions in the IHT system. In fact, they were removed from the language models when training the IHT engine, since we used bi-grams, and punctuation chars do not notably improve the predictions.

Table III
COLLECTED DATA. EACH ROW HOLDS THE METRICS FOR EACH PARTICIPANT, WHO IS DENOTED AS P##[†].

Metric:	Time				rWER				WSR				SUS	
System:	Manual		IHT		Manual		IHT		Manual		IHT		Manual	IHT
Page Order:	45 1	46 2	45 2	46 1	45 1	46 2	45 2	46 1	45 1	46 2	45 2	46 1	note [‡]	
P01	276.3	411.2	463.5	331.9	5.9	2.8	7.0	3.4	100.0	105.5	31.2	25.2	70.0	75.0
P02	514.6	234.1	394.5	243.4	2.4	6.4	4.7	3.2	100.4	92.1	33.7	25.9	62.5	77.5
P03	207.2	336.1	178.0	136.9	9.4	2.4	5.7	2.5	100.4	93.1	38.6	28.2	50.0	37.5
P04	419.6	153.2	250.3	201.0	1.6	18.8	8.5	2.0	100.0	100.4	25.6	28.8	77.5	70.0
P05	319.4	253.0	296.0	249.8	5.1	4.5	2.0	3.8	88.4	100.0	22.6	32.4	77.5	87.5
P06	320.6	218.0	370.4	249.2	3.7	12.8	5.9	5.7	100.4	100.4	22.2	38.2	62.5	97.5
P07	231.6	257.4	201.4	136.2	11.1	3.7	4.9	9.8	101.7	109.4	22.2	29.9	62.5	72.5
P08	461.4	174.2	436.1	315.2	7.8	7.2	10.6	8.6	100.0	86.4	32.4	36.6	52.5	42.5
P09	393.9	380.9	341.2	215.8	12.8	2.8	9.8	14.1	96.5	96.1	33.7	29.4	72.5	90.0
P10	347.8	310.1	224.1	178.7	20.5	11.5	15.3	13.1	100.4	95.4	20.5	33.7	80.0	65.0
P11	218.4	196.8	208.8	190.0	4.1	4.9	9.0	3.4	100.4	88.4	45.6	34.6	65.0	70.0
P12	297.9	189.9	350.4	79.3	4.2	10.2	3.4	2.4	81.6	97.1	30.7	38.2	95.0	62.5
P13	244.2	287.8	342.5	397.1	8.6	2.8	4.5	5.5	105.3	99.5	22.6	27.3	50.0	50.0

Order = 1 means that the user transcribed that page initially with the corresponding engine (inversely, order = 2 means that the user already transcribed the page with the other engine before.)

[†] In some cases decimals have been padded to better display cell values.

[‡] Users only ranked once each system via the online SUS questionnaire, for that reason there are only two SUS columns.

C. Correlation Analysis

We considered significant correlations when the Pearson Coefficient $|r| > 0.5$ ($r \in [-1, 1]$). Additionally, the Coefficient of Determination $r^2 \in [0, 1]$ allowed us to determine how certain would be a prediction from a given measure.

1) *Correlation between trials:* Overall, IHT is more stable than the manual engine for all measured variables (see Table IV). What is interesting is the consistency between trials for the IHT engine; no matter if a page has been seen once, IHT will behave approximately in the same way. However, if the user has seen a page previously the manual approach will result in less transcription time — although the user will need to write all the words, exposed thus to potentially more errors, and taking into account that it is not a realistic scenario (see subsection III-A).

Table IV
BETWEEN-TRIAL CORRELATION AND DETERMINATION COEFFICIENTS.

System	Time		rWER		WSR	
	r	r^2	r	r^2	r	r^2
Manual	-0.29	0.08	-0.22	0.04	-0.001	0
IHT	0.62	0.38	0.61	0.37	0.63	0.39

2) *Correlation between metrics:* We found a correlation between time and rWER in the manual engine when the page has not been seen once: $r = -0.615$ ($r^2 = 0.37$). For the same engine, rWER and WSR seem to have a relative influence on the user subjective ratings: $r = 0.45$ ($r^2 = 0.20$); $r = -0.75$ ($r^2 = 0.57$). This fact reinforced our initial hypothesis (see Introduction). For the IHT engine we did not observed relevant between-metrics correlations.

D. Limitations of the Study

There are a number of reasons why we were unfortunately not able to achieve statistically significant differences between the tested engines in some cases. First, the limited size of the user sample was a primary factor of influence. Taking also into account that users were not experts in transcribing ancient documents, a dispersed behavior was expected (i.e., some users were considerably faster/slower than others, see Table III and Figure 6). Second, the pages were really deteriorated, making more difficult the reading for the users. For that reason, there is a great difference between the first time that a user had to transcribe a page and the subsequent attempts. Third, most of the participants had never faced neither any of the implemented engines nor the web UI before the study, so it is expected a logical learning curve prior to using such systems in a daily basis. A simplified starting level would minimize this effect for the task; however we tried to select a scenario as close as possible to a realistic setting. Finally, the web interface was just a prototype, and it is well known that a careful design of the UI is a primary factor to tap the possibilities of the IHT technology. However, despite of the above mentioned limitations, there is a comprehensible tendency to choose the IHT paradigm over the manual system. Additionally, as observed, the probability of improvement of an IHT engine over manual transcription revealed that the interactive-predictive paradigm worked better for all users.

IV. CONCLUSIONS AND FUTURE WORK

The advantage of IHT over manual transcription seems clear, although it goes beyond the human effort reductions achieved. The proposed interactive approach constitutes a

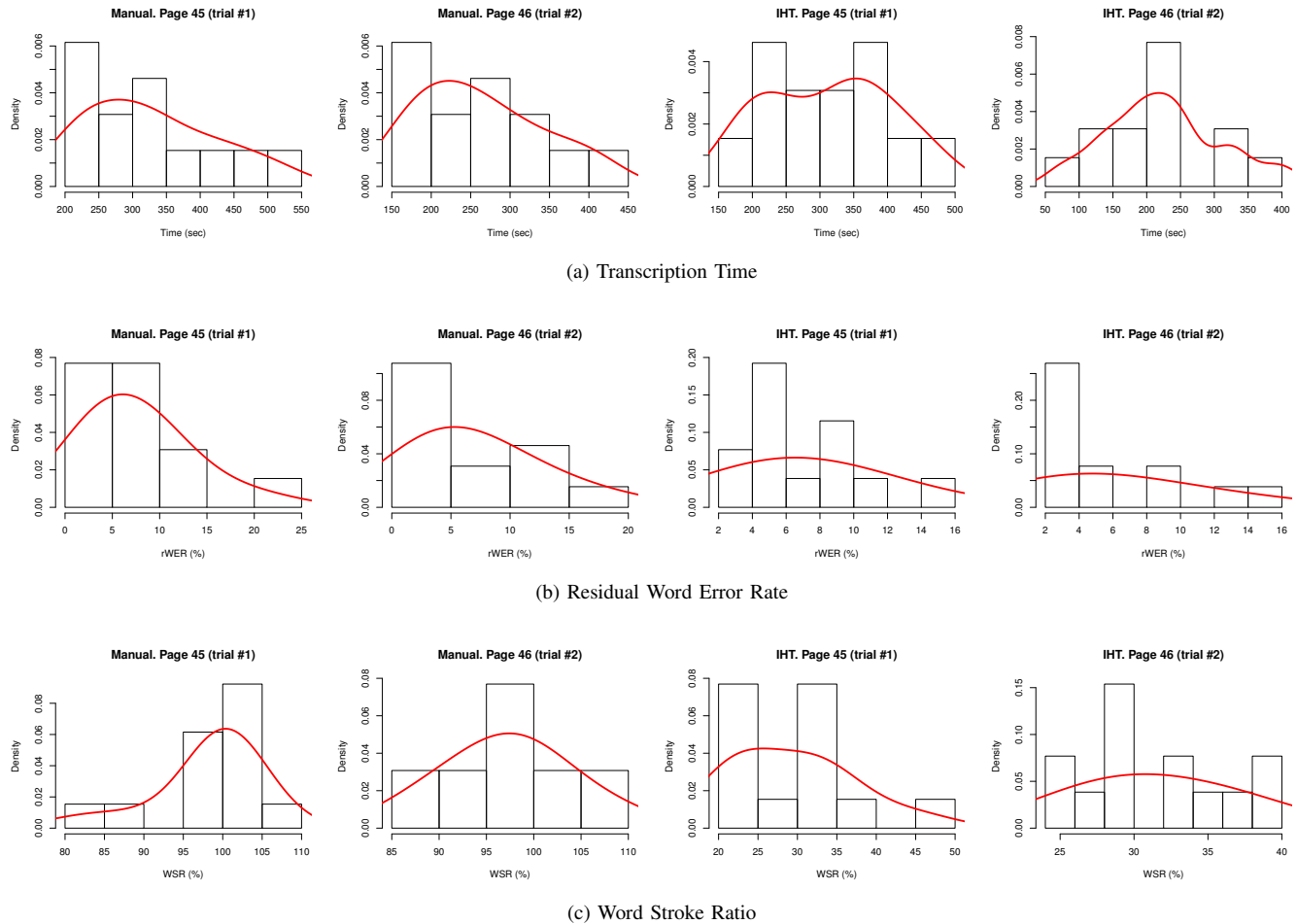


Figure 6. Probability distributions of measured performance metrics between users. Outer columns of this 4x3 matrix allow to compare both systems when transcribing a page for the first time (Vice versa for the inner columns.) Notice how gathered data could not be considered normally distributed in most cases.

much more natural way of producing correct text. With an adequate UI, IHT lets the users be dynamically in command: if predictions are not good enough, then the user simply keeps typing at her own pace; otherwise, she can accept (partial) predictions, thereby saving both thinking and typing effort.

Future work includes incorporating some of the modifications and enhancements that users insightfully reported at the end of our study. Thus, a much more extensive and large-scale evaluation of the IHT paradigm will be ready to be tested with professional transcribers.

ACKNOWLEDGEMENTS

We thank prof. José R. Navarro for his statistical advices, as well as the users who took part in the study. This work has been supported by the EC (FEDER/FSE), the Spanish MEC/MICINN under grant TIN2006-15694-C02-01, the research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018), and the MITTRAL (TIN2009-14633-C03-01) project.

REFERENCES

- [1] A. H. Toselli, V. Romero, M. Pastor, and E. Vidal, "Multi-modal interactive transcription of text images," *Pattern Recognition*, vol. 43, no. 5, pp. 1814–1825, 2009.
- [2] L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "reCAPTCHA: Human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.
- [3] E. Vidal, L. Rodríguez, F. Casacuberta, and I. García-Varea, "Interactive pattern recognition," in *Proc. of MLMI*, 2007, pp. 60–71.
- [4] A. Vinciarelli, S. Bengio, and H. Bunke, "Offline recognition of unconstrained handwritten texts using HMMs and statistical language models," *IEEE Trans. on PAMI*, vol. 26, no. 6, pp. 709–720, 2004.
- [5] M. Zimmermann, J. Chappelier, and H. Bunke, "Offline grammar-based recognition of handwritten sentences," *IEEE Trans. on PAMI*, vol. 28, no. 5, pp. 818–821, 2006.

- [6] V. Romero, A. H. Toselli, L. Rodríguez, and E. Vidal, "Computer assisted transcription for ancient text images," in *Proc. of ICIAR*, ser. LNCS. Springer-Verlag, 2007, vol. 4633, pp. 1182–1193.
- [7] V. Romero, A. H. Toselli, and E. Vidal, "Computer assisted transcription of text images: Results on the germana corpus and analysis of improvements needed for practical use," in *Proc. ICPR*, 2010, pp. 2017–2020.
- [8] A. H. Toselli, V. Romero, L. Rodríguez, and E. Vidal, "Computer assisted transcription of handwritten text," in *Proc. of ICDAR*, 2007, pp. 944–948.
- [9] V. Romero, A. H. Toselli, and E. Vidal, "Using mouse feedback in computer assisted transcription of handwritten text images," in *Proc. ICDAR*, 2009, pp. 96–100.
- [10] F. Drida, "Towards restoring historic documents degraded over time," in *Proc. of DIAL*, 2006, pp. 350–357.
- [11] V. Romero, L. A. Leiva, A. H. Toselli, and E. Vidal, "Interactive multimodal transcription of text images using a web-based demo system," in *Proc. of IUI*, 2009, pp. 477–478.
- [12] J. Brooke, "SUS: A "quick and dirty" usability scale," in *Usability Evaluation in Industry*. Taylor and Francis, 1996.
- [13] "R: A language and environment for statistical computing," <http://www.R-project.org>, R Foundation for Statistical Computing, 2009, ISBN 3-900051-07-0.
- [14] M. Stephens, "EDF statistics for goodness of fit and some comparisons," *Journal of the American Statistical Association*, vol. 69, no. 347, pp. 730–737, 1974.
- [15] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. of ICASSP*, 2004, pp. 409–12.