Word-Hunter: A Gamesourcing Experience to Validate the Transcription of Historical Manuscripts

Jialuo Chen, Pau Riba, Alicia Fornés, Joan Mas, Josep Lladós
Computer Vision Center - Computer Science Department
Universitat Autonoma de Barcelona, Spain
{jchen,priba,afornes,jmas,josep}@cvc.uab.es

Joana Maria Pujadas-Mora
Centre for Demographic Studies
Universitat Autonoma de Barcelona, Spain
jpujades@ced.uab.es

***Abstract*—Nowadays, there are still many handwritten historical documents in archives waiting to be transcribed and indexed. Since manual transcription is tedious and time consuming, the automatic transcription seems the path to follow. However, the performance of current handwriting recognition techniques is not perfect, so a manual validation is mandatory. Crowdsourcing is a good strategy for manual validation, however it is a tedious task. In this paper we analyze experiences based in gamification in order to propose and design a gamesourcing framework that increases the interest of users. Then, we describe and analyze our experience when validating the automatic transcription using the gamesourcing application. Moreover, thanks to the combination of clustering and handwriting recognition techniques, we can speed up the validation while maintaining the performance.**

***Index Terms*—Crowdsourcing; Gamification; Handwritten documents; Performance evaluation;**

I. Introduction

Despite the efforts in the last decades, the amount of historical manuscripts that have not yet been transcribed is still huge [3]. Consequently, they have not been properly indexed and their contents are not available through search platforms. Since manual transcription requires enormous human efforts, the challenge is to go towards an automatic transcription. However, the state of the art in handwriting recognition still need more development before trusting a completely automatic transcription. For this reason, human-assisted approaches based on handwriting recognition [9] and keyword spotting [14] are being used.

In the last years, the crowdsourcing strategy [18] has emerged as an interesting alternative. The key idea of crowdsourcing is to split the work in a big amount of micro-tasks and soliciting contributions from a large group of people, especially from the online community. Crowdsourcing at large has emerged as a novel business model in the social economy when users have a monetary incentive for their job. Platforms like Amazon Mechanical Turk [2] are now very popular. In the transcription of historical documents, crowdsourcing can be easily performed using specific applications [4], [15], web-based interfaces [17], [6], mobile applications [1], or even mobile applications with speech dictation [9]. In these scenarios, the contribution of users is generally voluntary, and their reward uses to be their satisfaction of collaborating in making the historical and cultural assets publicly available. Nevertheless, the transcription is still tedious, and many transcribers loose interest after a while. The challenge is how to keep human intelligence in the transcription process offering an engaging experience.

Gamification, defined as the application of game-design elements and principles in non-game contexts, has demonstrated to engage and keep the interest of users. Lately, it has been also applied to crowdsourcing activities [13], such as the $Digitalkoot$ [5] transcription games at $Facebook$. We believe that the transcription of historical document collections can be speed-up if we focus on two aspects: automatic transcription and manual validation through gamesourcing (understood as crowdsourcing via gamification). Firstly, when the automatic transcription is quite accurate, the time spent by the user to validate and correct errors is lower than manually transcribing from scratch, as demonstrated in [7].

In this paper we describe a gamesourcing experience to validate the results of a handwritten text recognition (HTR) system. First, we take advantage of the empirical observation that some words appear with high frequency. In demographic documents, that is the experimental corpus that we use, this fact is especially noticeable. Thus, after segmenting word images, a hierarchical clustering approach formulates predictions of repetitive words that can be transcribed by recognizing a small percentage of representatives. Two simple android games are designed and tested. One is addressed to prune clusters of outlier words. The second game is designed to validate the results of the HTR. Thanks to the combination of clustering and handwriting recognition techniques, we can avoid the validation of every single word. As a consequence, we can speed up the transcription while maintaining the performance.

In summary, our gamesourcing application, named *Word-Hunter*, is used to validate the automatic transcription, minimizing the human effort, and still engaging the users (even though the total time spent might be the same).

The rest of the paper is organized as follows. We describe the *Word-Hunter* gamesourcing application in Section II. The experiments are shown in Sections III and IV. Finally, Section V concludes the paper.

II. Word-Hunter: Gamesourcing Application

In this section we first describe the key crowdsourcing concepts. Afterwards we outline the system architecture and we describe the main components.

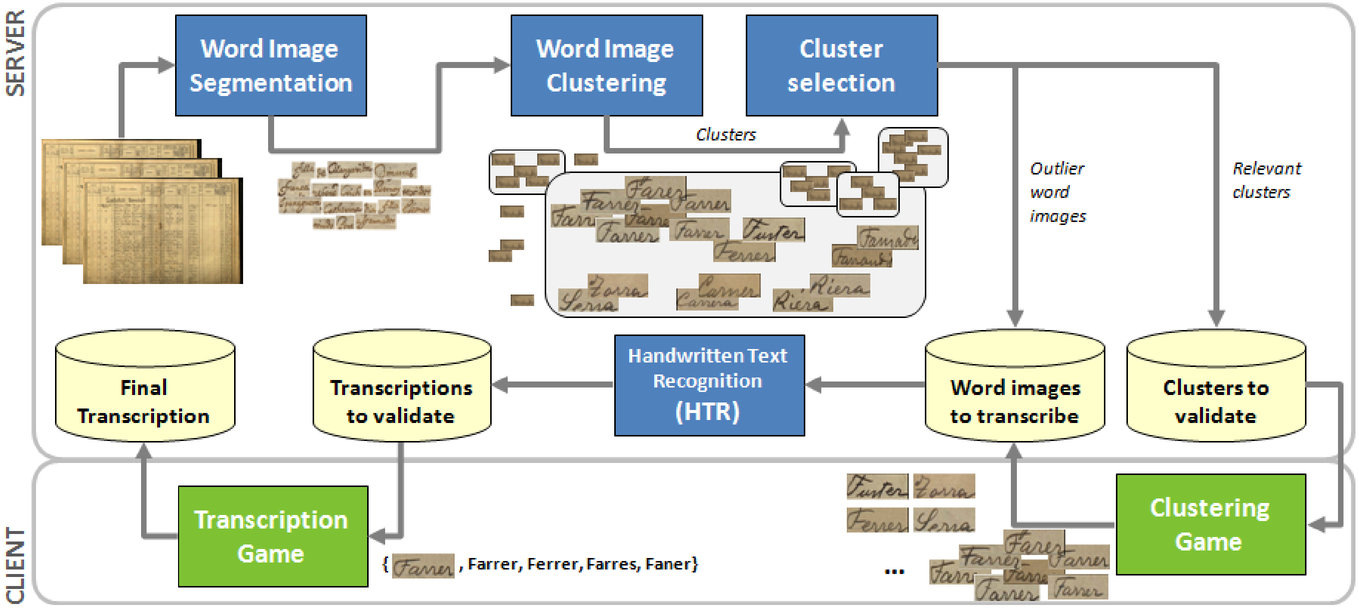

Fig. 1. System architecture. The server feeds the Android games with images, and analyzes the user's feedback in order to validate the transcriptions.

A. *Key aspects in crowdsourcing*

The principle of transcribing words using crowdsourcing was first introduced by reCAPTCHAs [12] (now a service of Google). It displays words taken from scanned texts that OCR programs cannot recognize. Hidden in the Turings test of CAPTCHAs of differentiating between humans and computers, human computations are implicitly performed to transcribe historical books word by word with a high performance close to 99%. The process of reCAPTCHA consists of requiring users to transcribe the word images that are displayed as a humanness test. The key strategy is to show to the user words with an unknown transcription, and *control* words, in which the transcription is known. If many users have a consensus when they are asked to type the un-transcribed words, the system considers that it is a valid transcription so it is automatically recorded. This basic principle is incorporated in gaming applications that generate transcriptions as a by-product of the engagement. For the sake of understandability of this paper, the key concepts that are involved in the gamification process are described below.

- Golden Tasks. It refers to the tasks associated to control words, i.e. words which transcription is known (ground truth or words already transcribed by OCR/HTR or humans with enough confidence). The purpose is to train the users at the beginning of the game, but also to check the user expertise, and hence their confidence. Usually, the number of golden tasks decreases as the game progresses.
- Validations. A validation is a user answer. A word is considered as validated when a minimum number of answers are obtained from different users. It is a parameter that can also depend on the expertise of users.
- Consensus. This parameter determines when a validated word is approved. One can opt for the majority vote or weighed vote according to users' expertise [11].
- Typology and expertise of users. In historical handwriting transcription, native users (that know the language) and experts (e.g. historians) may provide better answers.
- Engagement and Rewards. Instead of crowdsourcing with monetary incentives, gamification is a successful alternative. A key aspect in the design of the game is how to engage players to keep on playing (game score in terms of correct answers based on golden tasks, ranking, etc.)

B. *System Architecture*

The outline of the complete system architecture is shown in Figure 1. The main components of the system are hosted in a server, while the gamesourcing apps run in an Android client. Given a collection of handwritten documents to transcribe, snippets corresponding to word images are first extracted (in this paper we do not focus on the segmentation step, and the experimental work has been done with segmented word images). Word images are clustered to find high frequency words that can be jointly transcribed using a small percentage of representative instances. A quality assessment of clusters in terms of size and compacity aims to select clusters that are potentially valid, and to discard outlier word images (small clusters, isolated instances). These clusters are validated using the first of the proposed gamesourcing applications, the clustering game. It shows some instances of a given cluster, and the user is asked to confirm that those words are the same (so, these words really belong to the cluster).

The database of word images to transcribe contains isolated word images. These words correspond to words that do not belong to any cluster (discarded by the cluster selection module or by the clustering game app). On another hand, this database also contains words from validated clusters that will be transcribed together. The HTR module generates plausible transcriptions of these word images. The second gamesourcing app, the transcription game, is used to validate these transcriptions. In the following subsections we further describe the main components of the architecture.

C. Image Recognition Component

1) Word Clustering: The clustering algorithm aims to group word images that are visually similar, which means that they will probably have the same transcription. For this purpose, we use the dense SIFT descriptor and the k-means algorithm to perform a hierarchical clustering, following a tree structure. In the deeper levels of the clustering, the grid size of the SIFT descriptor increases (the SIFT feature vector is longer) so that we can highlight small differences between the words. The process stops when the clusters are compact.

2) Handwriting Recognition: The Handwritten Text Recognition (HTR) system is based on the adaptation of the Pyramidal Histogram of Characters (PHOC) attribute embedding to sequence learning. This approach is divided into two parts. The first stage corresponds to CNNs that embeds small windows of text (i.e. text patches) into the PHOC space. Then, this sequence of embeddings is recognized using BLSTM-RNNs. In this case, we do not use any dictionary or language model. For further details, the reader is referred to [16].

D. Word-Hunter Game

The *Word-Hunter* game consists of two sub-games.

1) Difference Game for Validation of Clusters: This game is designed to validate the word clustering algorithm. The system shows between 3 and 5 images belonging to the same cluster. If the cluster is correct, the player must select the option "They are the same". If there is an improperly clustered word, the user has to select it. If the cluster contains several outliers, the user must select "More than one are different", which means that the cluster is noisy and must be discarded. Figure 2 shows a screen-shot of this game.

The golden tasks (already known clusters) are used to teach the player and to give points for each correct answer. In the first levels, the amount of clusters and words is low, whereas in higher levels the amount of words and clusters significantly increase. Moreover, the player has to validate all words in each level within a limited time. If the countdown arrives to zero before completing the challenge, the player looses the game.

2) Match Game for Validation of Transcriptions: This game is designed to validate the output of the HTR algorithm. When the player selects one word, the system shows the most probable transcriptions (e.g. the n-best words) according to the HTR. The user must select the correct answer among these possibilities. If none of the transcriptions is correct, the user must press "None of these", so this word will be manually transcribed off-line. Figure 3 shows a screen-shot of this game.

As in the first game, players obtain points for each correct answer. Golden tasks (words with a known transcription) are used for learning, and for validation quality assessment. At higher levels, the amount of golden tasks decreases. Whenever the player selects a wrong answer, there is a penalty in the score. At each level, the player must validate the transcription of several words within a limited time. If the user can correctly validate all words within the given time, the player upgrades to the next level. Otherwise, the player looses the game.

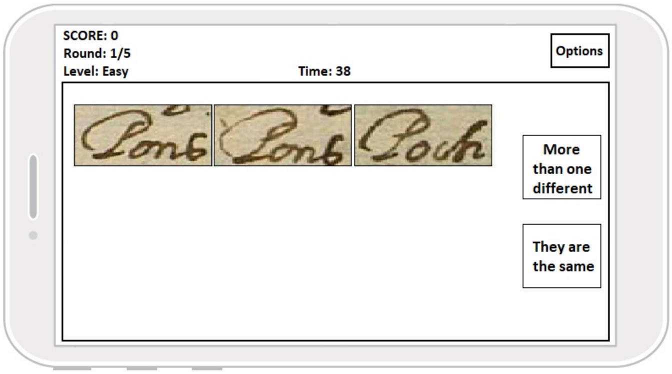

Fig. 2. Difference Game. The player has to validate the correctness of the cluster and remove possible outliers. In this example, "Poch" does not belong to this cluster "Pons", so the user has to select it.

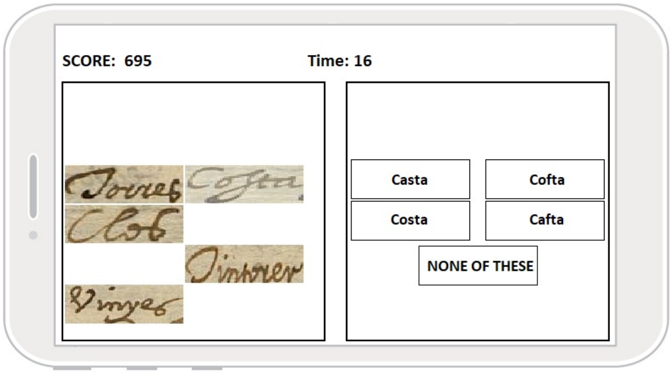

Fig. 3. Match Game. For the selected word (top right), its possible transcriptions are shown. The user should select the transcription "Costa".

III. Experimental Setup

This section describes the dataset and the typology of users that have been selected to participate in the experiment.

A. Dataset

Although our architecture is generic and any kind of documents could be used, we have chosen population documents. Population sources, such as marriage or census records, allow the study of the demographic behaviour and the understanding of the social and economic evolution of the past. In nominative sources, one of the most relevant keywords to index are the names and surnames. For the game, we have selected 938 instances of surnames from the marriage records of the Barcelona Cathedral [8]. From these, 250 word images are used as golden tasks, and the remaining 688 words for

TABLE I
TYPOLOGY OF THE 40 USERS THAT PARTICIPATE IN THE EXPERIMENT.

	Native / Foreigner	Men / Women	Standard / Experts
People	26 / 14	27 / 13	32 / 8

validation. The HTR has been trained with the training set of the ICDAR-IEHHR competition [8].

B. User Profiles Description

One of the important factors when analyzing the user feedback is the typology of the users taking part on the experiment, such as the gender, age, nationality or expertise. Table I summarizes the typology of the 40 users that participated in the experiment. We consider that there are 26 native users, because they know the language and the vocabulary (i.e. common surnames) appearing in these documents. The 8 experts belong to the fields of history and demography, with notions in paleography. This factor makes them more confident when reading historical manuscripts. Contrary, there are 14 foreigners that do not understand the language of the documents. In addition, most of them have a completely different mother tongue and language script, such as Chinese, Hindi, Arabic or Persian.

C. Tasks Description

We have asked the users to play to both games.

1) Task 1. Difference Game: The players use the first game to validate the words within each cluster and remove any outliers. The golden tasks are groups of images of already known clusters. At first, the number of golden tasks is 4, and progressively decreases in the next levels. Contrary, the number of clusters to validate increases in the next levels.

2) Task 2. Match Game: The players use the second game to validate the transcription provided by the HTR system. Here, the golden tasks are words with a known transcription to teach users. Here the number of golden tasks is 10 in the first levels, and progressively decreases to 2 in the next levels.

IV. RESULTS AND DISCUSSION

In this section we show and analyze the players' learning curves and their validation accuracy, and also, the improvement of the use of gamesourcing for validation tasks.

A. Learning Curve

For analyzing the learning curve and the time spent at each level, we monitor the users' behaviour through the analysis of the golden tasks. Figures 4 and 5 show the learning curves in terms of the average number of errors and the time spent in each level of the game. As expected, the amount of errors and time required for each level decreases as long as the user keeps on playing. However, when users play for long time, the amount of errors slightly increases again, maybe because more words must be validated within the same time, and player has to answer quick.Some users have reported that they experienced eyestrain that can explain this decrease on their performance.

We also observe that native users (plots in red and dark blue color) usually make less mistakes and require less time to complete each level. Moreover, in the ”Match Game”, the frustration of foreigners when validating transcriptions is so high that none of them has played up to level 11.

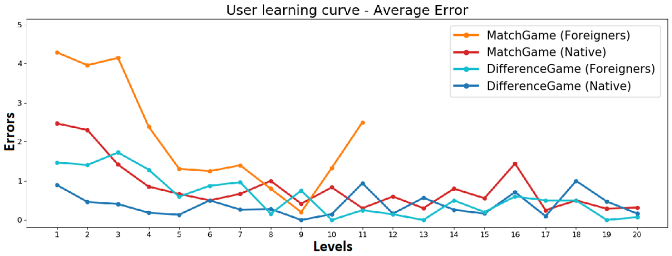

Fig. 4. Learning curve - Average Errors.

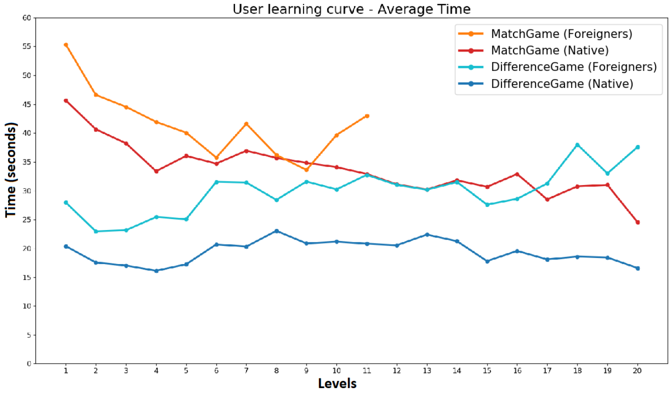

Fig. 5. Learning curve - Average Time spent per level.

B. User's Validation Accuracy

Before validating the performance of the users, one must decide which is the minimum amount of validations required for each word, and the percentage of consensus that is required for considering a word as validated. Figure 6 shows the variations in the percentage of words that would be correctly validated, wrongly validated, and also the percentage of words without any agreement when we consider different percentages of required consensus. We observe that, as expected, a higher consensus ensures a higher validation accuracy, but the total amount of validated words decreases. Thus, the objective is to find a suitable trade-off, trying to minimize the errors while maximizing the amount of validated words. We have tested with different values, and we have observed that a good trade-off is to obtain a minimum of 5 validations for each word, and force a consensus of minimum 65%. This means that a word is considered as validated if more than 5 people have validated this word, and more than 65% of them agree in the answer.

Once we have set these values, we have analyzed whether the different type of users cause differences in the performance. In this sense, we have not experienced any difference

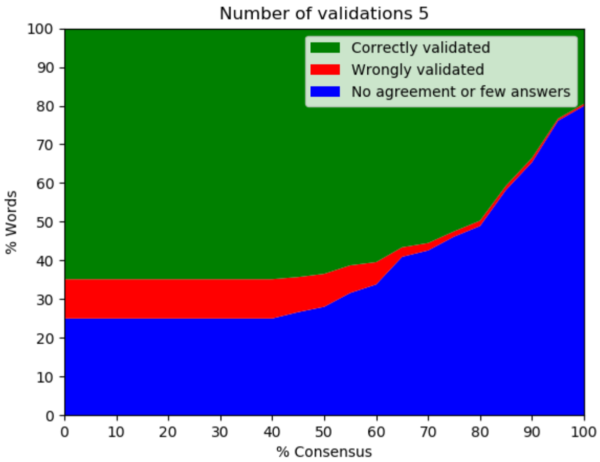

Fig. 6. Variations of the percentage of correctly validated, wrongly validated and no agreement in terms of the variation of the percentage of consensus, when words have a minimum of 5 validations.

in terms of gender. However, there is an important difference between foreigners, native and expert users. Table II shows the validation performance of foreigners, natives and experts when validating the transcription of words using the ”Match Game”. We first observe that foreigners usually disagree (there is low consensus). If we decrease the minimum amount of validations to 3, we could assume that the 35% of words can be validated. However, from these validated words, the Word Error Rate (WER) is close to 40%.

In the case of native users, the consensus is higher, and the percentage of validated words ranges from 61% to 69% if we consider 3 or 5 minimum validations. In both cases, the WER is around 9-10%. Finally, in the case of experts, we have observed that the consensus is lower, although the word error rate (WER) is very low. Notice that, when only 3 validations are required, the 43% of words can be considered as validated, and from them, the WER is 2.5%.

TABLE II
RESULTS WITH A MINIMUM CONSENSUS OF 65%. ALL VALUES ARE BETWEEN 0-100%.

	Foreigner		**Natives**		**Experts**	
# Validations	3	5	3	5	3	5
Validated	35	8.75	69.5	61.19	43	20.93
- Correct (100-WER)	60.7	71.43	89.7	91.22	97.3	94.0
- Wrong (WER)	39.3	28.57	10.3	8.78	2.7	6.0
No consensus	12.5	7.5	8.96	8.06	12.79	9.3
Few responses	52.5	83.75	21.49	30.75	44.19	69.77

C. Performance Analysis of the Game

Finally, we analyze whether the transcription could be speed up by transcribing clusters instead of individual words. For example, if one cluster contains many ’Smith’, we could assume that all the words belonging to that cluster can be labelled as ’Smith’. Consequently, the user does not need to validate each individual word within each cluster (only a certain percentage is required), so we can save human efforts.

For the evaluation, we have designed three scenarios:

Scenario 1: Users do not play. In this case, we suppose that the user is not present (there is no game). Thus, the validation is completely automatic, as follows: For each cluster, we check if the HTR system has provided the same transcription for most words belonging to this cluster. If there is a consensus, all words within the cluster are labelled using the same transcription. Otherwise, this cluster is not considered as validated, so these words must to be transcribed one by one.

Scenario 2: Users play the first game. Here, users validate the clusters and remove any incorrect words (outliers) using the ’Difference Game’. The removed words will have to be individually transcribed and validated. Given that the transcriptions are not validated by the users, we follow the same procedure as in the first scenario: if there is a consensus in the transcriptions of words within a cluster, then all these words are labelled with the same transcription.

Scenario 3: Users play the two games. Users validate the clusters and the transcriptions. In this case, after validating the clusters, if there is a consensus in the transcriptions of these clusters, then all these words are labelled with the same transcription.

TABLE III
RESULTS FOR THE DIFFERENT SCENARIOS: USERS PLAY AT NONE, ONE OR TWO GAMES. ALL VALUES ARE BETWEEN 0-100%.

	No Game		**Game 1**	**Game 1&2**	
Consensus	>65%	>90%	>65%	>65%	>90%
Validated words	61.4	7.4	35.7	15.3	13.8
- Correct (100-WER)	74	100	90.7	98.1	100
- Wrong (WER)	26	0	9.3	1.9	0
Pending words	38.6	92.6	64.3	84.7	86.2
Human Effort	None	None	Medium	High	High

The results obtained in each one of the scenarios are shown in table III. In the first scenario we observe that, if the consensus is minimum 65% (i.e. more than the 65% of words in a cluster have the same transcription), the 61.4% of words in the collection are automatically validated, so only the 38.6% of words remain as pending to be individually transcribed and validated. However, note that in this case, only the 74% of transcriptions are correct. If we force a higher consensus (above 90%), we can only consider that the 7.4% of all the words in the collection are validated, but we can ensure a word accuracy of 100% (WER=0). Note that in this case, we could validate a small amount of words, but in a completely automatic manner (without human intervention).

In the second scenario, the human effort is medium (users only play one game). Here, we observe that the 35,7% of words are validated, and from these, the 90.7% are correct (WER=9.3%). Note that the word accuracy is not perfect because users can also make errors when validating the clusters.

In the third scenario, the human effort is high because users must play the two games to correct the clusters and transcriptions. In this case, the consensus of each cluster must take into account the transcription provided by the HTR, and also, the transcriptions corrected by the users. If the consensus is above 65%, the word accuracy is 98.1%. Contrary, if we force a higher consensus (above 90%), the 13.8% of words are correctly validated (WER=0). Note, however, that in this latter case, the percentage of pending words to be individually validated also increase (86%).

In the third scenario we have noticed that, if users validate the transcription of at least the 40% of the words within each cluster, the final results are exactly the same (the WER does not increase). This means that we could save a significant human effort, because the 'Match Game' (for validating transcriptions) is much slower than the 'Difference Game'.

D. Discussion

From the analysis of the results and the users' feedback, we can draw several conclusions that would guide the design of improved games. First, the typology of users is a key factor. As expected, the knowledge of the language, vocabulary and the handwriting style indeed benefits the validation accuracy. This means that native users are more trustworthy, so less validations are necessary to validate the transcriptions.

Second, frustration is also an important factor. In the "Match Game", when the difficulty of the transcriptions increases, the amount of errors made by foreigners raises, so they get frustrated and stop playing. However, foreigners have an almost comparable performance when validating clusters, because it is just necessary to check the visual similarity of words, instead of reading them. Therefore a game based on shape or visual similarity could be used worldwide.

Third, the number of golden tasks, the percentage of consensus and the number of required validations should be dynamic, adapted to the expertise of users. Indeed, since the learning curve shows peaks at higher levels, the system should increase the amount of golden tasks until the player demonstrates a high performance again. Indeed, the answers of the users to golden tasks could be used to infer their reliability or expertise level. Thus, players that usually answer the correct option are more trusty in their validations.

Forth, words should be classified according to their difficulty. As stated in [10], the user should read a difficult word within their context (previous and subsequent words) before deciding its transcription. In our case, the word difficulty could be set according to the HTR confidence and/or the difficulties of users in reaching a consensus. In those cases, these words should only appear at higher levels. Also, a higher amount of validations should be required. In fact, once those difficult words are correctly validated, they should be used as golden tasks to teach users how to avoid incorrect transcriptions.

Last, the game should not penalize because of time, because at higher levels, there are more words to validate. Therefore, players are forced to answer quickly, so they are more prone to errors.

V. Conclusion

In this paper we have described our experience when validating the automatic transcription of historical documents using a gamesourcing application. We believe that the lessons learned could help to improve future gamesourcing applications, and also, to define more suitable trade-offs between user engagement and validation performance. Moreover, thanks to the combination of clustering and handwriting recognition techniques, we can avoid the validation of every single word. Thus, we can speed up the validation of the transcriptions while maintaining the performance.

Acknowledgment

We thank all users that participated in the experiments. This work has been partially supported by the Spanish project TIN2015-70924-C2-2-R, the Ramon y Cajal Fellowship RYC-2014-16831, the CERCA Programme / Generalitat de Catalunya, and RecerCaixa (XARXES, 2016ACUP-00008), a research program from Obra Social "La Caixa" with the collaboration of the ACUP.

References

[1] A. Amato, A. Sappa, A. Fornés, F. Lumbreras, and J. Lladós, "Divide and conquer: Atomizing and parallelizing a task in a mobile crowdsourcing platform," in *CrowdMM*, 2013, pp. 21–22.

[2] Amazon-Mechanical-Turk. Url:https://www.mturk.com/.

[3] V. Bachi, A. Fresa, C. Pierotti, and C. Prandoni, "The digitization age: Mass culture is quality culture. challenges for cultural heritage and society," in *EuroMed*, vol. 8740. Springer, 2014, pp. 786–801.

[4] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Aletheia-an advanced document layout and text ground-truthing system for production environments," in *ICDAR*. IEEE, 2011, pp. 48–52.

[5] DigitalKoot. Url:http://www.digitalkoot.fi/.

[6] A. Fornés, J. Lladós, J. Mas, J. M. Pujades, and A. Cabré, "A bimodal crowdsourcing platform for demographic historical manuscripts," in *DATeCH*. ACM, 2014, pp. 103–108.

[7] A. Fornés, B. Megyesi, and J. Mas, "Transcription of encoded manuscripts with image processing techniques," in *Digital Humanities*, 2017.

[8] A. Fornés, V. Romero, A. Baró, J. I. Toledo, J. A. Sanchez, E. Vidal, and J. Lladós, "Competition on information extraction in historical handwritten records," in *ICDAR*, 2017, pp. 1389–1394.

[9] E. Granell, V. Romero, and C. D. MartnezHinarejos, "Multimodality, interactivity, and crowdsourcing for document transcription," *Computational Intelligence*, 2018.

[10] E. D. Karnin, E. Walach, and T. Drory, "Crowdsourcing in the document processing practice," in *ICWE*. Springer, 2010, pp. 408–411.

[11] R. Kern, H. Thies, C. Bauer, and G. Satzger, "Quality assurance for human-based electronic services: A decision matrix for choosing the right approach," in *ICWE*. Springer, 2010, pp. 421–424.

[12] C. M. D. A. L. Von Ahn, B. Maurer and M. Blum, "Recaptcha: Human-based character recognition via web security measures," *Science*, no. 5895, p. 14651468, 2008.

[13] B. Morschheuser, J. Hamari, and J. Koivisto, "Gamification in crowdsourcing: a review," in *HICSS*. IEEE, 2016, pp. 4375–4384.

[14] A. Santoro, C. De Stefano, and A. Marcelli, "Assisted transcription of historical documents by keyword spotting: a performance model," in *ICDAR*, 2017, pp. 971–976.

[15] E. Saund, J. Lin, and P. Sarkar, "Pixlabeler: User interface for pixel-level labeling of elements in document images," in *ICDAR*. IEEE, 2009, pp. 646–650.

[16] J. I. Toledo, S. Dey, A. Fornés, and J. Lladós, "Handwriting recognition by attribute embedding and recurrent neural networks," in *ICDAR*, 2017, pp. 1038–1043.

[17] Transkribus. Url:https://transkribus.eu/.

[18] M.-C. Yuen, I. King, and K.-S. Leung, "A survey of crowdsourcing systems," in *PASSAT and Socialcom*. IEEE, 2011, pp. 766–773.