https://doi.org/10.52326/ic-ecco.2021/CS.03

Cultural Heritage Digitisation Using Game with a Purpose

Maria Luca[1], Victoria Bobicev[2]
[1,2] Technical University of Moldova, St. cel Mare bvd, 168, Chisinau, Republic of Moldova,
codreanu.maria@iis.utm.md, victoria.bobicev@ia.utm.md, https://utm.md/

***Abstract* — The digitisation of cultural heritage is a priority area in the European Union. This task require a close collaboration between researchers in humanities and computer scientists. The paper presents our effort to create a a training set of recognized handwritten text. To this mean we propose an online game with a purpose (GWAP). This is a human-based computing technique in which some stages of a computational process are transmitted to humans in a form of an online game. The input of the game is a line of handwritten text and the user have to type in its transcription. We investigated how to make this in an attractive and enthralling way in order to involve more users and obtain a large set of transcribed texts.**

Keywords* — *cultural heritage digitisation, automate manuscript recognition, training set, games with a purpose

I. Introduction

The digitisation of cultural heritage is a priority area on the European Union's digital agenda. The Commission Recommendation on digitisation, online accessibility and digital preservation of cultural material [1] represents a milestone in digital cultural policy. The Recommendation invites the States to pool their resources and involve the private sector in digitising cultural material, in order to increase online accessibility of European cultural heritage, boost engagement of citizens and growth in Europe's creative industries. Furthermore, digitised material should be made accessible through the European cultural platform providing access to a wide array of digital content from Europe's libraries, archives and museums.

Huge collections of old documents stored in the museums and libraries cannot be made easily available for the large public as they need to be preserved in special conditions of temperature, humidity and light. Today, there is a solution for this problem. Plenty of old documents are already published by online digital libraries around the world for preservation and for making them available online for everyone who is interested [8].

Our republic is also working towards digitisation of our cultural treasures, in order to facilitate its preservation, online access and re-use. The decision no. 478 of July 4, 2012 on the National Program for computerization of the sphere of culture for the years 2012-2020, published in the Official Gazette [2] described in details the plans of our cultural heritage digitisation and its online access.

Multiple cultural organizations have been involved in this process including museums, libraries, research centries on history and arts, everyone who is working with valuable historical objects. This work, however cannot be done without computer scientisits, who is able to perform the necessary operations and organize online access to the digitised content.

We present our effort to create a training set of recognized handwritten text. To this mean we propose an online game with a purpose (GWAP). This is a human-based computing technique in which some stages of a computational process are transmitted to humans in a form of an online game (this process is called gamification) [7].

II. Digitisation of Our Cultural Heritage

The digitisation of the cultural heritage is necessary for two major reasons:

1. Historical cultural objets preservation. All old objects have been gradually destroying, especially if they are exposed for the large public.
2. Free access for everyone who is interested. Digital copy of the historical documents uploaded online are easily accessible by anyone to explore without additional damage to fragile originals.

The term "digitisation" (we use British spelling, in USA "digitization" is used) is defined in the Collins dictionary as "turning information into a form that can be read easily by a computer" or "transformation of data into a digital form so that it can be directly processed by a computer" [1] and this is only the first step for digital

[1] https://www.collinsdictionary.com/dictionary/english/digitize

https://doi.org/10.52326/ic-ecco.2021/CS.03

preservation. Digital preservation is to maintain something over a long period of time in digital form [10]. However, there is the other goal of the digital preservation activities, namely, organization of the free access to the digital copies of the historical objects. This is especially important for the historical documents: old books, letters, manuscripts and other texts; they are of great importance for the researchers in the fields of history and arts.

The Institute of Mathematics and Informatics of the Moldavian Academy of Science took part in the research project "Information technologies and resources for digitizing the Romanian historical-literary heritage from the 17th-20th centuries printed in Cyrillic alphabet" within the State Program "Development of e-Infrastructure data in the field of research, development and innovation in the Republic of Moldova" , the years 2018-2019 [9]. The aim of the project was the development of an instrument to support the digitisation and transliteration of texts printed in Romanian with Cyrillic characters. In their report they listed the main steps of the digitisation process for these documents:

- Obtaining the image by scanning of the original physical documents from the library, using the software provided by the scanner. The desired quality is 600 DPI (Dots Per Inch) or more.
- Preparing images for OCR (Optical Character Recognition) with a software, which performs massive automatic corrections of image defects, for example, corrects the tilt angle of the page or cleans some small spots. This work is quite laborious when processing large volumes of text.
- Optical recognition of the Cyrillic characters of the scanned texts.
- Transliteration of the results obtained in modern Latin script.
- Manual or automated processing of the result obtained for the final correction of the text.

Thus, the whole process cannot be completed without manual verification of the final text even for the printed sources and we worked with handwritten texts. At the moment, handwritten tetxs are still a great problem for the automate systems. In order to create systems able to read handwritten text we need a large volume of training material that have to be created manually.

III. Preprocessing of the Old Texts

In our work, we used a limited set of manuscripts already scanned and published on the online portal "National Digital Treasure" of the Republic of Moldova, www.digi.emoldova.org

Scanned old documents are the result of the first step in the process of their digitisation; in addition to the steps mentioned in the previous section, there is one more important step: image segmentation for the further optical character recognition and its post-processing [3]. In comparison to printed texts, the process of manuscripts' recognition is considerably more difficult. The causes of the difficulties are:

- The text is written on various types of paper, which has lost its quality over time;
- All manuscripts have different writing style, depending on the period in which they were written;
- The use of different alphabets, such as Cyrillic, Latin and so called alphabets of transition [13];
- The texts are written disorderly, various signs being placed above and below the line;
- Abbreviations and other special marks.

Various machine learning algorithms are used for this task including recent neural network architectures. Even a small set of training data can significantly improve the model's performance on corresponding test data [4].

At the text segmentation stage the whole text on the page is split in rows for further recognition. This process is described in [5]. Then, the rows are presented to online users for recognition in a fun way. The answers, introduced by users are stored along with the original handwritten row forming the labeled instances for training a machine learning model [6].

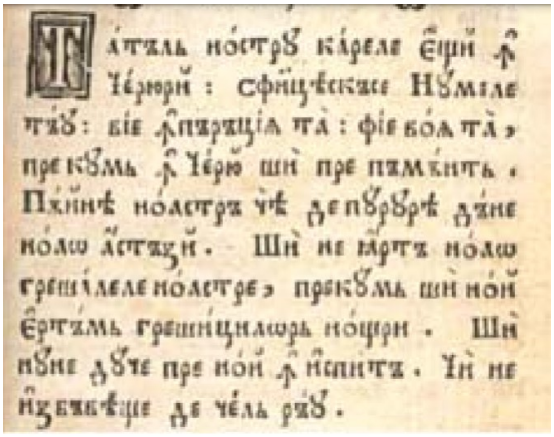

Figure 1. An example of old printed text. [3].

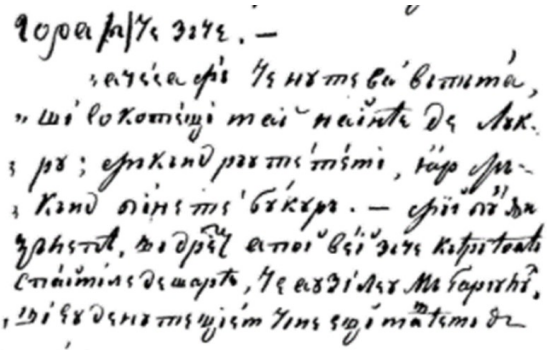

Figure 2. An example of old handwritten text. [5]

Figure 1 presens an example of old printed text and Figure 2 contains an example of old handwritten text.

https://doi.org/10.52326/ic-ecco.2021/CS.03

Both texts are old and difficult to read and understand by a modern person, but printed text is seemingly easier.

Before the text can be recognized it should be segmented in rows and words. This task is complicated by differen heght of the letters and various symbols that appear above and below the letters of the text: it is seen on the picture on figures 1 and 2.

Figure 3 presents a process of text segmentation described in [5]. The output of this step are separate rows of text that should be recognized.

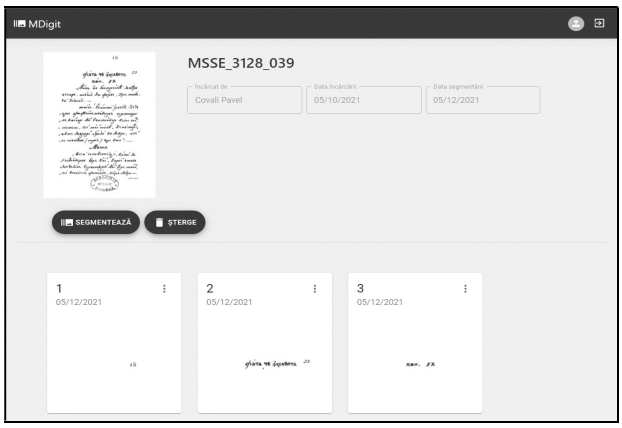

Figure 3. An example text mage segmentation. The whole page with text is above and the pages with extracted rows of the text are below. [5]

IV. Using the Game with a Purpose

Even after segmentation in rows the handwritten text is extremely difficult to recognize. The modern machine learning methods can be trained to perform this task with some acceptable accuracy but they need a training set to be trained. A training set is a set of examples with the answers for the given task. The bigger is the training set the better is the model of a machine learning method, the better are the results of text recognition. For the task of handwritten text recognition, a training set is a volume of handwritten text fragments with their transcriptions in printed words. These transcriptions are made manually by the specialists. However, there are not many such specialists and they are not able to make a set of transcriptions large enough for the training. We decided to organize crowdsourcing of such transcriptions.

The word "crowdsourcing" is a combination of "crowd" and "outsourcing" and it is a sourcing model in which individuals or organizations obtain goods or services including ideas, voting, micro-tasks, and finances from a large, relatively open, and often rapidly evolving group of participants. It typically involves using the internet to attract and divide work between participants to achieve a cumulative result [11].

In our case we planned to create a web interface in which web users can transcribe the rows of handwritten texts adding pieces of data to the training set. In order to attract more users to the transcription process we use game with a purpose (GWAP).

A human-based computation game or game with a purpose is a human-based computation technique of outsourcing steps within a computational process to humans in an entertaining way (gamification) [14]. The tasks presented in these games are usually trivial for humans, but difficult for computers. One of the first games with a purpose was ESP Game, also known as Google Image Labeler "an online game in which players label images with words that describe them" [14]. In this game people provided meaningful and accurate tags for web images as a side effect of the game; for example, an image of a man and a dog was labeled 'dog', 'man' and 'pet'. The game was fast, enjoyable and competitive.

It is well known that there are a large number of web users who play online games. If a game is fun, there is a good chance that enough online users will play it [12].

The Entertainment Software Association reported that "over 200 million hours are spent every day playing computer and video games in the United States." Indeed, by the age of 21, the average American had spent more than 10,000 hours playing such games, the equivalent of five years of full-time work 40 hours a week.

People play not because they are interested in solving a computational problem, but because they want to have fun. The GWAP approach is characterized by three motivating factors: an increasing proportion of the world's population has access to the Internet; certain tasks are impossible for computers, but easy for people; and people spend a lot of time playing computer games.

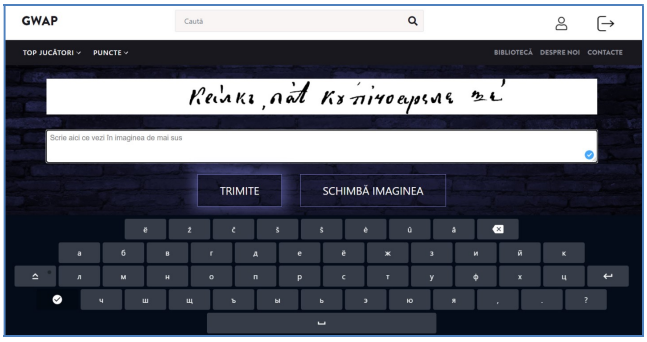

Figure 4. An example of the game interface. [6]

One of the most important aspects of GWAP is the ability to be enjoyable. The fact that people enjoy the game makes them want to continue playing, producing in turn more useful results. Setting time limits for game sessions introduces the challenge into a game in the form of a timed response. The timed response is effective for introducing the challenge as long as the goal is not trivial for the players.

One of the most direct ways to motivate players is to award points for each successful exit produced during the game. For the ESP game, pairs of players receive points for each image for which they successfully agree on a

https://doi.org/10.52326/ic-ecco.2021/CS.03

word (which then becomes a label for the image). Using points increases motivation by providing a clear connection between effort in the game, performance (obtaining the winning condition) and results (points). A summary of the score after each game also provides players with performance feedback, making it easier to assess progress on score-related goals (such as exceeding a previous game score and completing all task instances within the set time limit).

The designed game is a Game with Exit Agreement similar to ESP game.

Initial configuration: Two users are randomly chosen by the game itself from all potential players.

Rules: In each round, both are given the same input and they must produce outputs based on the input. The game instructions indicate that players should try to produce the same result as their partners. Players cannot see each other's results and cannot communicate with each other.

Winning condition: Both players produce the same output; they are not synchronized in time, but they have to produce a resilt at some point while the input is displayed on the screen.

The application is written in PHP, uses JQuery, its database is MySQL. The database contains data about users, games and images.

The main modules of the application are:

- Enrollment and authentication of the user;
- Visualisation of the user account information;
- Game rounds;
- Assignment of ranks and motivation points for players; visualization of the ranking.

The game interface presented on Figure 4 contains a row of handwritten text to be recognized, a input field where the player should introduce the response and a virtual keyboard with the necessary letters as the letters used in the handwritten documents are specific and some of them cannot be found on the standard keyboard.

The main purpose of the game is to produce a large amount of useful and clean data. Each of these three objectives ("large", "clean" and "useful") has important implications for the design of the proposed game. First, to collect large amounts of data, the game must be attractive to users. The next requirement is that the data be clean. First, players must be able to produce high quality annotations and the game should encourage users to enter relevant data. We award points as a motivating factor, but this can cause players to enter irrelevant data.

There may not always be enough players available online to match a human player with another human player. Therefore, an important part of designing an online game is building a bot that can work instead of a player.

V. Conclusion and Future Work

The presented work is a part of the project dedicated to our cultural heritage digitisation, namely old manuscripts written in transitional alphabet. Scanned manuscripts have to be transcribed and transliterated in modern romanian. Our task is to create a large set of manually transcribed handwritten examples in order to train a machine learning metod. Game with a purpose (GWAP) is a good way to solve this problem with the help of users which playing online will transcribe the handwritten texts. In the paper, we present a game developed for this purpose.

Our future plans include uploading the game on the web, attracting users to this game, collecting a set of transcribed rows of handwritten texts and experimenting with machine learning models trained on this set of data.

References

[1] EUROPEAN COMISSION Commission Recommendation of 27 October 2011 on the digitisation and online accessibility of cultural material and digital preservation [online]. ELI: http://data.europa.eu/eli/reco/2011/711/oj, 2011. [Accessed: 08.08.2019]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32011H0711

[2] Hotărârea nr. 478 din 04 iulie 2012 cu privire la Programul național de informatizare a sferei culturii pentru anii 2012-2020, *Monitorul Oficial Nr. 143-148 art. 531*.

[3] E. Boian, C. Ciubotaru, S. Cojocaru, A. Colesnicov, L. Mahlov. "Digitizarea și recunoașterea și conservarea patrimoniului cultural – istoric", *AKADEMOS, nr. 1 (32),* pp. 61-68, martie 2014.

[4] J. Walker, Y. Fujii, A. C. Popat. "A Web-Based OCR Service for Documents" *13th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 21-22, april 2018.

[5] P. Covali, "Automatizarea segmentarii imaginii pentru digitizatrea patrimoniului cultural", *Master Thesis. UTM, Moldova,* 2021.

[6] M. Luca, "Recunoașterea textelor vechi pentru digitizarea patrimoniului cultural", *Master Thesis. UTM, Moldova*, 2021.

[7] L. von Ahn, L. Dabbish. "Designing Games with a Purpose". *Communications of the ACM, 2008, Vol. 51 No. 8*, pp. 58-67, 2008.

[8] A. Juan, V. Romero, J. A. Sánchez, N. Serrano, A. H. Toselli, E. Vidal. "Handwritten Text Recognition for Ancient Documents." *First Workshop on Applications of Pattern Analysis, PMLR 11*, pp. 58-65, 2010.

[9] S. Cojocaru, A. Colesnicov, L. Malahova, T. Bumbu, Ș. Ungur, "Raport științific final privind executarea proiectului de cercetări științifice aplicative „Tehnologii și resurse informaționale pentru digitizarea patrimoniului românesc istorico-literar din secolele 17-20 tipărit cu alfabet chirilic", *Vladimir Andrunachievici Institute of Mathematics and Computer Science*, pages 24, 2019.

[10] P. Caplan, "What is Digital Preservation?". *Library Technology Reports. 44 (2): 7*, 2008. Retrieved 2016-10-26.

[11] E. Estellés-Arolas, F. González-Ladrón-de-Guevara, "Towards an Integrated Crowdsourcing Definition", *Journal of Information Science, 38 (2)*, 2012, pp. 189–200.

[12] D. Vickrey, A. Bronzan, W. Choi, A. Kumar, J. Turner-Maier, A. Wang, D. Koller, "Online Word Games for Semantic Data Collection." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 533-542.

[13] S. Cojocaru, C. Ciubotaru, A. Colesnicov, L. Malahov, T. Bumbu, "Instrumentar pentru digitizarea și transliterarea textelor tipărite în limba română cu caractere chirilice." *REVISTA BIBLIOTECII ACADEMIEI ROMÂNE, Anul 2, Nr. 2,* iulie-decembrie 2017, p. 27-38.

[14] L. von Ahn, "Games With A Purpose". *Computer:* 2006. pp 96–98.