Global Journal of Computer Science and Technology: E Network, Web & Security
Volume 22 Issue 1 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Website Text Translation and Image Translation from a URL using Optical Character Recognition (OCR)

By A.H.M. Saiful Islam, Eshita Agnes Purification, Fahima Akter Anni & Kishor K. Baroi

Notre Dame University

Abstract- Now-a-days we are almost completely dependent on information system for our day-to-day work. Almost every organization of different sectors has their own website. These websites are visited not only by the native people but also by the foreigners. But sometimes they are unable to do so because of language barrier. At present, many translating tools are available but they are either for translating text of a website or translating text from an image. At some cases people have to copy the text and then translate it separately which is a lot of hassle and time consuming. We aim to implement a website translator which will take the URL of any website and translate it in any language. It can also translate the text of the images of that website. We have also created some more new algorithms for URL translation, English to Bangla number translation and English to Arabic number translation.

GJCST-E Classification: I.7.5

WEBSITETEXTTRANSLATIONANDIMAGETRANSLATIONFROMAURLUSINGOPTICALCHARACTERRECOGNITIONOCR

Strictly as per the compliance and regulations of:

Website Text Translation and Image Translation from a URL using Optical Character Recognition (OCR)

A.H.M. Saiful Islam [α], Eshita Agnes Purification [σ], Fahima Akter Anni [ρ] & Kishor K. Baroi [ω]

Abstract- Now-a-days we are almost completely dependent on information system for our day-to-day work. Almost every organization of different sectors has their own website. These websites are visited not only by the native people but also by the foreigners. But sometimes they are unable to do so because of language barrier. At present, many translating tools are available but they are either for translating text of a website or translating text from an image. At some cases people have to copy the text and then translate it separately which is a lot of hassle and time consuming. We aim to implement a website translator which will take the URL of any website and translate it in any language. It can also translate the text of the images of that website. We have also created some more new algorithms for URL translation, English to Bangla number translation and English to Arabic number translation.

I. Introduction

We live in a world of information system at present. We are very dependent to various websites for information about almost everything. For this purpose, people all over the world goes through numerous websites every day. But all websites are not available in their native languages. Around 75% of the world's population does not speak in English according to BBC - UK report1[2]. Here comes the need for translating the contents of the websites. Also, sometimes the images of the websites contain texts which are also need to be translated.

Google translator is widely used for this translation purpose. It can translate texts of any websites using the url of the website. But it doesn't translate the texts inside the images of that website. If anyone searches for an educational website and there is an image of a notice, he/she will not be able to read it as it won't be translated using google translator.

For image translation there are also many apps and websites which are widely used to translate the texts inside of an image to any desired language. But they only deal with images. OCR (Optical Character Recognition) is widely used for the image translation method. It is a technology that recognizes text within a digital image [4]. It is commonly used to recognize text in scanned documents and images [4].

In our work, we tried to create a platform where the users will be able to translate the whole website in any language using the URL and they will also be able to translate the image texts too.

We have also created a platform which will convert random images where the numbers will also be translated from English to any languages. We worked with only Bangla and Arabic numbers here. But English numbers can also be translated to other language numbers too only by editing the algorithm we created.

We organized this paper in this way:

Section

1. Gives the introduction of our work, section.
2. Eplains the implementation details of our website, section.
3. Includes the three algorithms we created, section.
4. Presents the outcomes of the experiments, as well as a comparison to the current procedures and the last section.
5. Contains the conclusion and future work.

II. Implementation Details

In our work, we create three types to translate such as website URL translation, image URL translation and image file (.png or .jpg) translation. Firstly, for website URL translation, we take a website URL as input to call that website from google website and run in our website along with a translation tool to translate that website in any language we want and we can see as figure 1.

Author α: Notre Dame University. e-mail: saiful@ndub.edu.bd

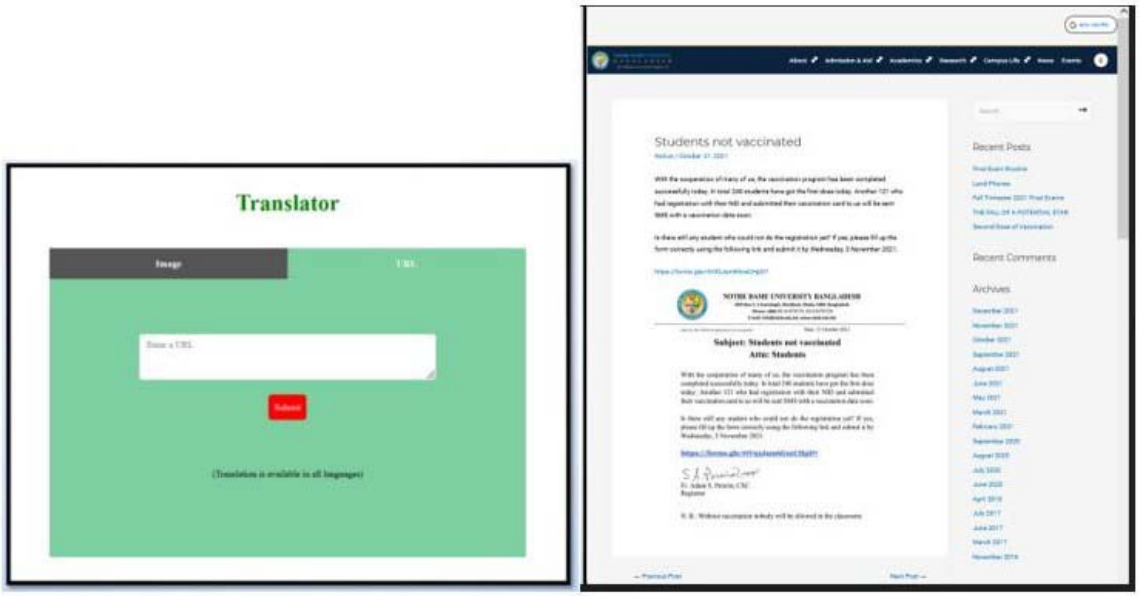
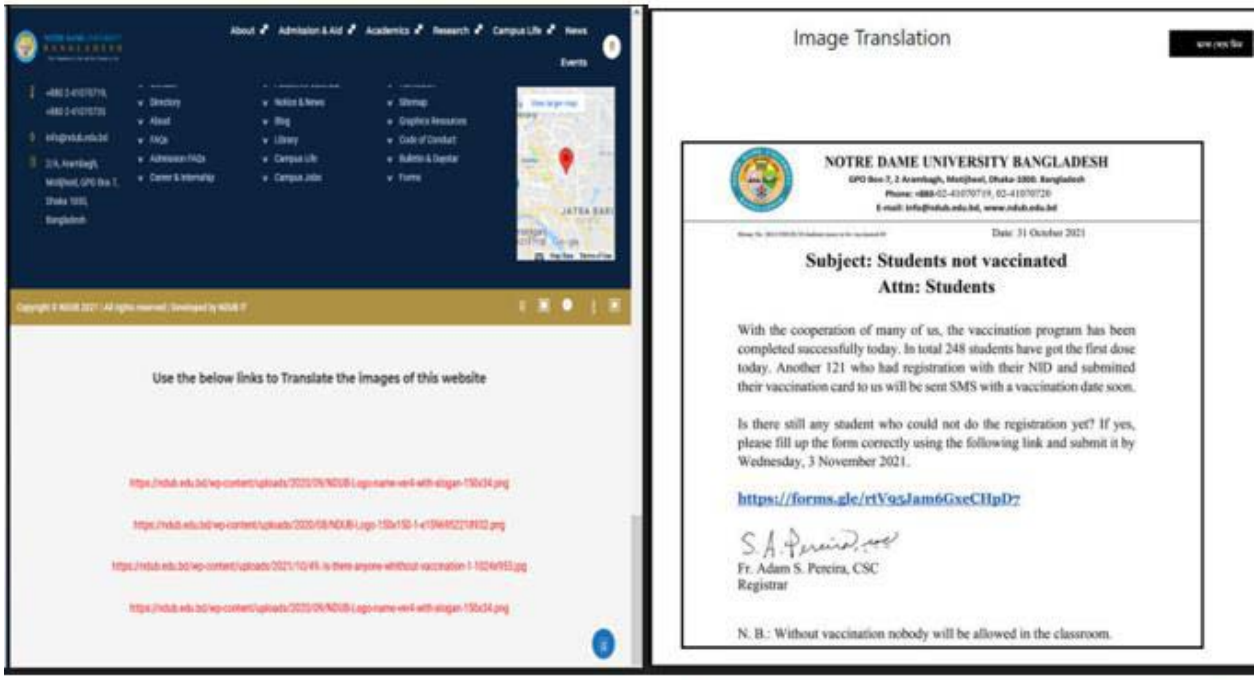


Figure 1: The homepage of the website and the website url to be tested

Secondly, for image URL translation, when we put a website URL and run that in our website it also collects all the image URL that website has and show it at the end of our website. By selecting an image URL, we collect the image from google and convert it to word by using tesseract OCR and save it in a file. After that we call the .txt file and show it to our website. When we select an image URL, we see that .txt file along with the image and translation tool. Now we can translate the image and read the image text in any language. The figure 2 portrays the translation process of the text from an image url. We also show the image text in a format so that it is easy to understand and easy to read.



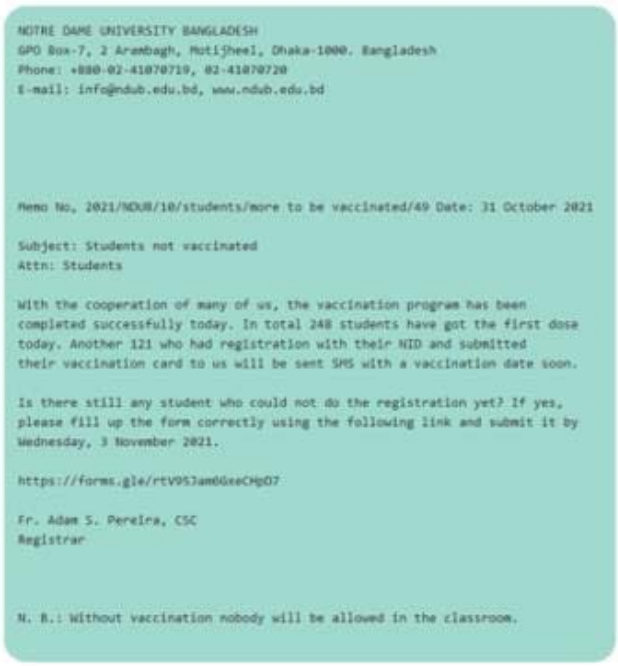


Figure 2: Translation of the text from the image from a website

Lastly, for image translation, we take an image as an input to translate the image text along with numbers. We use our own algorithm to translate the image text and numbers as a sample we use English to Bengali or English to Arabic/Persian Language translation. When we put an image, it converts the image to text and put it in a .txt file. Then using a function to convert the numbers from English to Bengali or Arabic/Persian numbers and show the whole file in our website after converting the numbers and we get the following results in figure 3. And figure 4 shows the translated view of an image sample from English to Bengali and from English to Arabic respectively.

Figure 3: View of the image insert module of the website

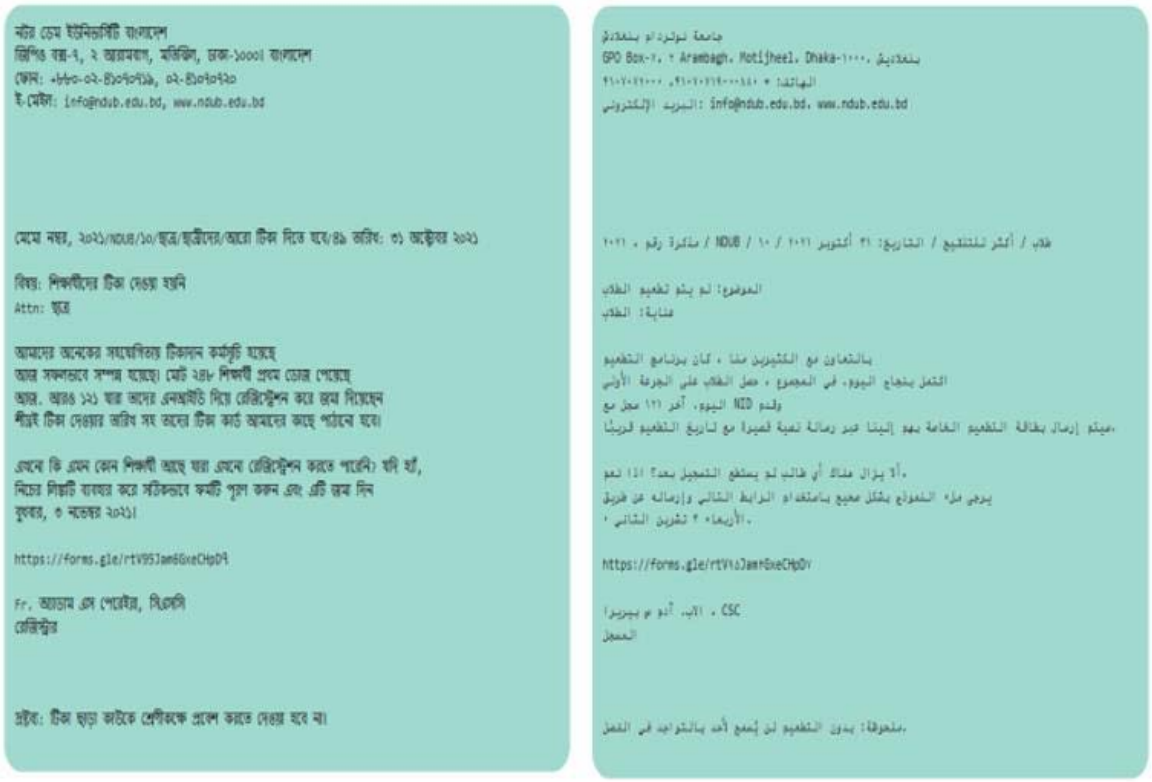


Figure 4: Translated view of the image in Bangla and Persian from English

III. Algorithms

In this section our own algorithms are discussed. We have implemented these three algorithms in our website. Figure shows the first algorithm which is used for translating texts of a random image from English to Bangla.

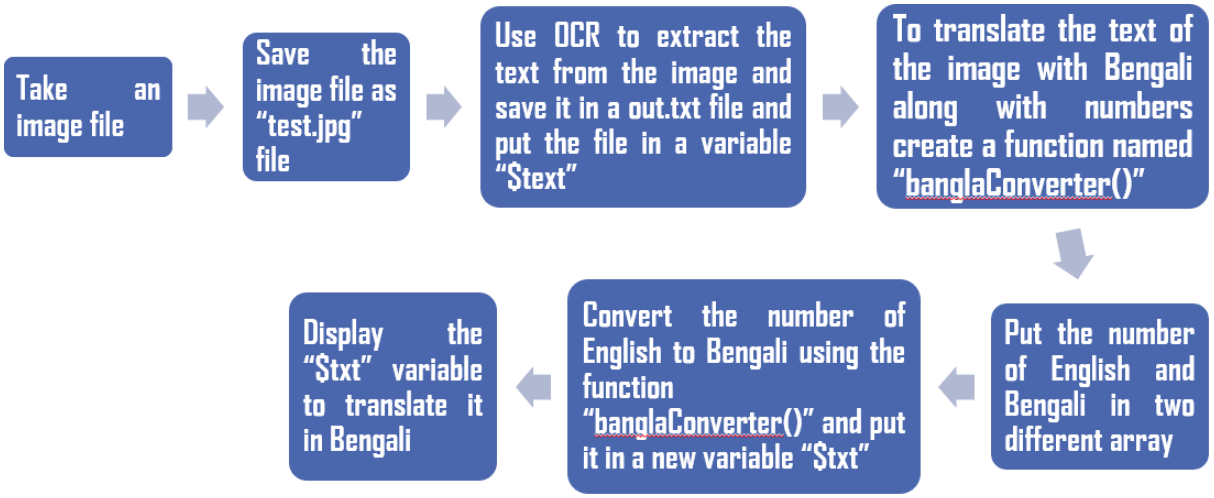


Figure 5: Algorithm to Image Translation (English to Bengali)

Figure 6 describes the algorithm to translate the text of a random image from English to Persian.

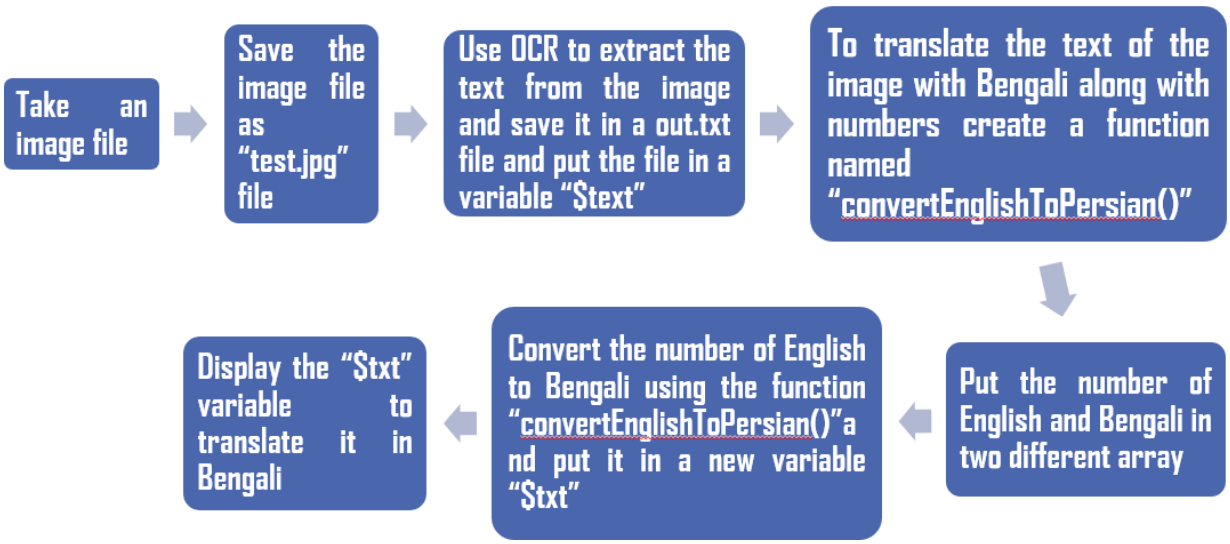


Figure 6: Algorithm to Image Translation (English to Persian)

Figure 7 describes the algorithm to translate the text of an image from a website url from English to any language.

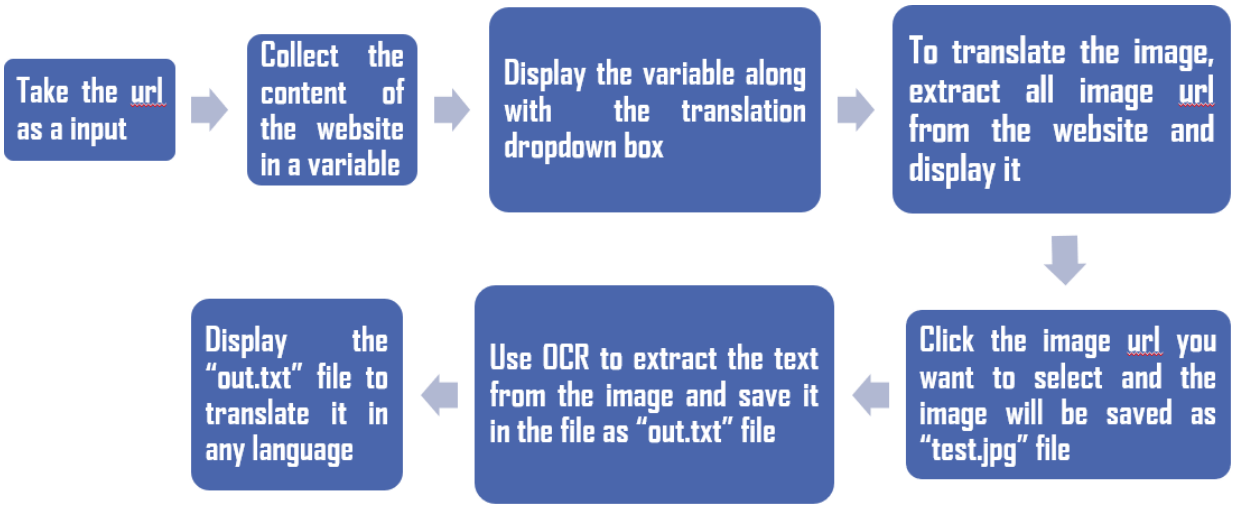


Figure 7: Algorithm for URL Translation

IV. Results Analysis

This section describes the results of our works. In all three sectors of our work, we used the term Accuracy to calculate the performance of them.

Accuracy: It is defined as the ratio of translated words and total words. Here w is the number of properly translated words, and W is the number of total words.

$$\text{Accuracy} = w / W$$

We have experimented Up to 70 websites and up to 50 random images with our approach. Our website reaches almost 83 percent accuracy in the field of website translation and 85 percent in the field of translation of images from those websites. The accuracy of the translation of the random images from English to Bangla reaches 98 percent and from English to Persian it reaches almost 93 percent. We tested this approach with various text fonts, and our website accurately translated them all.

V. Conclusion and Future Work

We have implemented an easier and user-friendly website which takes an url as input and translate the website in any desired language. Using this platform, users will be able to get the information of any website in their comfortable language and it is also time-saving as it translates any website using only a URL. It also translates the texts inside the images of the website. The users are also able to extract the texts of a random image and we have used our own algorithm to translate the English numbers to Bangla and Arabic numbers.

In future, this paper will be helpful to build a mobile application where one can add camera module to take an image and translate it through the app where they can translate numbers too. This paper can also help to build an app or a website that will be able to take any url from any barcode and translate both the text and images. In future this paper will be helpful to create a new algorithm to translate text of all the images of the website in a single webpage along with the web text just like the original website.

References Références Referencias

1. Shruthi Kubatur, Suhas Sreehari, Rajeshwari Hegde "An Image Processing Approach to Linguistic Translation", Dept. of Electrical & Computer Engg, University of Windsor, Windsor, Canada Dept. of Telecommunication Engg, B.M.S. College of Engineering, Bangalore, India, December 2011.
2. Rijwan Khan, Aryan Kaushal, Ayush Agarwal, Avdhesh Kumar "Tourist's Translator based on Digital Image Processing and Hybrid Translation", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-9 Issue-5, March 2020.
3. G. R. Hemalakshmi, M. Sakthimanimala, J. Salai Ani Muthu "Extraction of Text from an Image and its Language Translation Using OCR", International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), ISSN: 2394-2320, Volume-4 Issue-4, April 2017.
4. https://www.necc.mass.edu/wp-content/uploads/accessible-media-necc/uncategorized/resources/What-is-OCR.pdf.
5. Azmi Can Özgen, Mandana Fasounaki, Hazim Kemal Ekenel, "Text detection in natural and computer-generated images" 2017.
6. R. Smith, "An overview of the Tesseract OCR engine", Document Analysis and Recognition 2007. ICDAR 2007. Ninth International Conference on, vol. 2, pp. 629- 633, 2007, September.
7. A. Canedo-Rodriguez, S. Kim, J. H. Kim, and Y. Blanco-Fernandez, "English to Spanish translation of signboard images from a mobile phone camera," IEEE Southeastcon 2009, Atlanta, GA, 2009, pp. 356-361. DOI: 10.1109/SECON.2009.5174105.
8. Seethalakshmi R., Sreeranjani T.R., Balachandar T., Abnikant Singh, Markandey Singh, Ritwaj Ratan, and Sarvesh Kumar, "Optical Character Recognition for printed Tamil text using Unicode," Journal of Zhejiang University SCIENCE, ISSN 1009-3095, 2005.
9. S.K, Vijaya Kumar, et al., "FLD based Unconstrained Handwritten Kannada Character Recognition," International Journal of Database Theory and Application, December 2000.
10. Pratik Madhukar Manwatkar, Dr. Kavita R. Singh, "A technical review on text recognition from images," IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO), 2015.