Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 158–167

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Augmented Reality for Scene Text Recognition, Visualization and Reading to Assist Visually Impaired People

Imene OUALI[a], Mohamed BEN HALIMA[a], Ali WALI[a]

[a]*REGIM-Lab: Research Groups in Intelligent Machines, National Engineering School of Sfax, University of Sfax, Tunisia*

Abstract

Reading traffic signs while driving a car for visually impaired people and people with visual problems is a very difficult task for them. This task is encountered every day, sometimes incorrect reading of traffic signs can lead to very serious results. In particular, the Arabic language is very difficult, making recognizing and viewing Arabic text a difficult task. In this context, we are looking for an effective solution to remove errors and results that can sometimes end someone's life. This article aims to correctly read traffic signs with Arabic text using augmented reality technology. Our system is composed of three modules. The first is text detection and recognition. The second is Text visualization. The third is Text to speech methods conversion. With this system, the user can have two different results. The first result is visual with much-improved text and enhancement. The second result is sound, he can hear the text aloud. This system is very applicable and effective for daily life. To assess the effectiveness of our work, we offer a survey to a group of visually impaired people to give their opinion on the use of our application. The results have been good for most people.

Keywords: Text Visualization, Text detection, Text recognition, Natural Scene, Augmented Reality, VGG19.

1. Introduction

Visually impaired people encounter several problems every day when reading text. These problems make their lives very difficult. In this article, we aim to make these people as independent as possible. Current works have limitations that make them impractical and unreliable. Moreover, these methods have not yet been commercialized. In this context, we propose an Android application based on augmented reality (AR) to offer more interactivity to users. The objective of the design and development of this system is to produce a mobile application to help people who are visually impaired or who have difficulty reading texts. By adding, to obtain clearer and improved textual information.

E-mail address: Imene3ouali@gmail.com

10.1016/j.procs.2022.09.048

In addition to facilitating navigation and understanding of their environment. The application has a simple and easy-to-use GUI. When AR technology is applied, the app adds some objects (displays digitally processed non-real text) to the real world. AR is the integration and addition of the user's environment with real-time digital information. In this article, we study the importance of using AR as an optimal solution to obtain a text. The legibility of the text is sometimes difficult to read. There are several issues, such as text style, color, environmental conditions, target surface shape, etc. These issues are poorly addressed in the literature. Our goal is to manage these issues well. A recommended solution is the use of AR technology. The use of AR using smartphones will add value to greater mobility. We propose a system that consists of three axes; text detection and recognition, text visualization, and finally, text-to-speech conversion. The results provided to the user are the conversion of text to speech and the visualization of Arabic text in real-time with smartphones. First, a detection module is applied to this captured image to detect the text that is in the captured image. Then, a text recognition tool is applied to this detected text image. Then, the recognized text is displayed on the screen of the user's smartphone. The user can interact with the content viewed. The information is provided as a 2D image. Thus, the user will have clear and readable text. Finally, visually impaired people can listen to the text with audio assistance with sound. However, this research emphasizes the process of reading and viewing text. This system must be able to arouse the interest of users, and applies in real life. The benefit of using advanced hardware is a better AR-enabled smartphone app that visually impaired people can experience for themselves.

Our proposed system is developed with the Unity application development system with the Vuforia extension. The latter allows the use of text images as input. These text images can be viewed in 2-dimensional (2D) animations. Plus, you can listen to it out loud. The Vuforia extension makes it possible to transform into simple and inexpensive digital experiences. The system assists the user with graphical and audio assistance. This application will allow peoples to better understand the world around them.

The remainder of this paper is organized as follows: Section 2 defines several related works. Section 3, presents the importance of augmented reality technology. Section 4, presents the proposed approach. Section 5 presents the evaluation and experimental test. Finally, Section 6 is reserved for the conclusion and future work.

2. Related Work

In this section, we first present some methods of text detection and recognition. These methods attempt to find relevant solutions. Secondly, we present some text visualization methods based on augmented reality. These methods aim to find effective solutions to give a good visualization to the users. Thirdly, we present some methods for converting text into speech. These methods aim to find effective solutions to give a good visualization to the users.

2.1. Text detection and recognition methods

In this subsection, a comparative study of different methods of text detection and recognition of natural images that are based on AR is presented. The following table (Table 1) gives a brief overview.

Table 1: Comparison between relevant related work Approach of text detection and recognition methods based DL

Ref	Year	Methods	Testing accuracy
[1]	2021	CNN	97%
[2]	2021	CNN + DBLSTM + DBLSTM + SVM	98.72%
[3]	2021	SVM + HOG + EHD	85%
[4]	2021	CNN + YOLOv3	96%
[5]	2021	CNN + SVM + XGBoost	96,3%
[6]	2021	CNN	95%
[7]	2021	CNN	70%
[8]	2021	DCNN	99.95%
[9]	2021	CNN	97.22%

A CNN-based handwritten Arabic character recognition model is reported on [1]. The results of this method attained accuracy values of 97%. Furthermore, Yahia et al. [2] propose a method for multilingual online handwriting recognition based on SVM networks and hybrid bidirectional long-term memory (DBLSTM). Moreover, Mamoun et al. [3] offers a text handwriting recognition system based on Edge Histogram Descriptor (EHD), Oriented Gradient Histogram (HOG), Feature Extraction and Support Vector Machine (SVM) as a classifier HOG and EHD gives optimum characteristics of Arabic handwritten texts by extracting directional properties from the text. In addition, Imran et al. [4] present an architecture to have a mechanism for detecting and recognizing license plates. This 53-layer CNN-based architecture is based on the YOLOv3 object detection algorithm variant. Furthermore, Naseem et al. [5] develops a system for recognizing and developing Arabic handwritten characters, developing a conventional neural network (CNN) with a hybrid model using Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost classifiers). In addition, Mohammad et al. [6] propose a CNN-based architecture for the recognition of Arabic and Latin handwritten alphanumeric characters. This architecture also recognizes the characters of license plates. Furthermore, Yahia et al. [7], offers new methods of data augmentation that generate more shapes and dynamic variations to improve the performance of recognition systems. Moreover, Rami et al. [8] present a CNN-based contextual model for recognizing offline handwritten Arabic text, characters, numbers, and isolated words. Furthermore, Mridul et al. [9] offer a DL-based system called LWSINet: LightWeight Script Identification Network (6-layer CNN). This system makes it possible to identify video scripts.

2.2. Text visualization methods

In this subsection, a comparative study of the different text visualization methods that are based on AR is presented. The following table (Table 2) gives a brief overview.

Table 2: Comparison between relevant related work Approach of text visualization method

Ref	Year	Methods	Language of text	Visualization tool
[10]	2022	Unity 3D engine	Arabic	Smartphone
[11]	2021	Vuforia	Arabic	Mobile
[12]	2021	JSON	German-English	Smartphone
[13]	2021	Vuforia	Bangladesh	Smartphone
[14]	2020	Unreal Engine 4s	Italy	Smartphone
[15]	2020	Vuforia	English	Mobile
[16]	2019	ARx2	Arabic language, English language	Smartphone
[17]	2019	Vuforia	Chinese	Web
[18]	2018	Vuforia	Aceh language, Indonesian language	Smartphone

Imene et al. [10] presents an AR-based architecture to visualize Arabic texts with diacritics in real-time using the Unity 3D engine. In addition, Imene et al. [11] presents an AR-based architecture to visualize Arabic letters with diacritics in real-time using Vuforia. Furthermore, Shane et al. [12] offer TeMoTopic, a visualization component for the temporal exploration of subjects in text corpora. Besides, Hosain et al. [13] presents a method to visualize characters with an interactive 3D view using a mobile camera-based AR application. In addition, Mori, et al [14] offers a virtual interactive experience created for the photo gallery of Jesi (Italy). Moreover, Imene et al. [15] presents an AR-based text recognition architecture to help the visually impaired. Besides, Henda et al. [16] develops an AR mobile application for real-time Arabic text translation. This application consists of three main components, which are: text detection, text extraction, and text translation. Besides, Yunqiang et al. [17] presents a natural feature recognition algorithm. The work is based on an inverted BOVW (Bag of View Words) and WebAR (Web-based AR) model. In addition, Zalfie Ardian et al. [18] develops a mobile application called ARgot. ARgot is a real-time translation application from Aceh to the Indonesian language.

2.3. Text to speech methods

In this subsection, a comparative study between different methods of text to speech is presented. The following table (Table 3) gives a brief overview.

Table 3: Comparison between relevant related work of text to speech

Ref	Year	Methods	Language
[19]	2021	ASR	Mandarin
[20]	2021	BERT	Multi Languages
[21]	2021	DenoiSpeech	English
[22]	2021	RNN	Chinese
[23]	2021	ASR	English
[24]	2021	Transformer-TTS, CycleVAE-VC et un vocodeur LPCNet	Japanese
[25]	2021	Extracts speaker and language embeddings from acoustic characteristics	English and Japanese
[26]	2021	G2P	Vietnamese
[27]	2021	LightTTS	Chinese and English
[28]	2021	NN-KoG2P	Korean
[29]	2021	Tacotron2	Chinese
[30]	2021	VC	English
[31]	2021	BERT	Mandarin
[32]	2021	Emotional multi-speaker TTS	Chinese
[33]	2021	PPG	Kestrel

Wang et al. [19] proposes data selection strategies on augmented TTS data, and the efficiency of the synthesized data can be significantly improved for ASR modeling of children. Furthermore, Liping et al. [20] presents a BERT model of speech to extract prosody information embedded in speech segments to improve the prosody of synthesized speech in neural speech synthesis (TTS). Besides, Zhang et al. [21] develops DenoiSpeech, a TTS system that can synthesize clean speech for a speaker with noisy voice data. Moreover, Feng-Long et al. [22] offers a hybrid TTS based on trajectory paving to improve its synthesis performance. A combination of the Transformer encoder and RNN-based decoder architecture where the two-level linguistic representation, both at the word and letter level of the Chinese phonetic alphabet, is exploited to generate a coherent trajectory and fluid speech settings. Furthermore, Changfeng et al. [23] presents a method for pre-training transformer-based encoder-decoder-based Automatic Speech Recognition (ASR) models using sufficient text in the target domain. Besides, Keisuke et al. [24] present this method to generate highly intelligible speech that preserves the individuality of dysarthric speakers by combining Transformer-TTS, CycleVAE-VC, and an LPCNet vocoder. Moreover, Detai et al. [25] propose a method to obtain unraveled speaker and language representations through mutual information minimization and domain adaptation for multilingual text-to-speech (TTS) synthesis. Besides, Dang-Khoa et al. [26] offer an app for the Vietnamese language. The dictionaries were then used to train different types of grapheme-to-phoneme converters (G2P). Moreover, Song et al. [27] offer a lightweight multi-speaker multilingual text-to-speech system, named LightTTS, which can quickly synthesize Chinese, English, or code-switched speech from multiple speakers in a non-autoregressive fashion using a single model. Besides, Hwa-Yeon et al. [28] propose a new Korean G2P model architecture, reflecting the characteristics of Korean pronunciation, called Korean G2P based on a neural network (NN-KoG2P). In addition, Cheng et al. [29] extend Tacotron2 with a height prediction task to capture discrete representations related to height. Moreover, Xinyuan et al. [30] present a framework for forming a non-parallel many-to-many voice conversion (VC) model based on the encoder-decoder architecture. Besides, Zilong et al. [31] provide a universal BERT-based model that can be used for various tasks in the Mandarin front-end without changing its architecture. Moreover, Junjie et al. [32] offer a chapter-based comprehension system for Chinese novels, to automatically predict speaker and emotion tags based

on the context at the chapter level. In addition, Weiwei et al. [33] treats TN text normalization as a neural machine translation problem and presents a purely data-driven TN system using the Transformer framework.

3. Importance of augmented reality technology

Augmented reality is a technology that integrates information from the real world with information from the virtual world. This technology is interactive in real-time and can operate in two or three dimensions. Thanks to this functionality of merging the real world and the virtual world. The user can improve their understanding and interaction with the real world and the virtual world at the same time. AR technology can be widely applied to various fields such as healthcare, education, visually impaired assistance, etc. AR systems are portable and can be used anywhere, such as on mobile devices, tablets, smart glasses, mobile augmented reality, etc. As a result, this development has become an important research direction of AR technology in mobile computing. The goal of the AR system is to improve user interactions and increase real-world perceptions by complementing 2D virtual objects with the real world that appear to co-exist in the same place in the real world. AR is a technology used in computer systems and in modeling the real world to support and assist human activities. Through the use of AR technology, users are in virtual space and real life. AR technology enables 2D display. In AR, several engines are available. For example, Vuforia is mentioned. The latter uses real-time computer vision technology to identify and track 2D target images. Vuforia SDK supports target image types, 2D. Vuforia's application programming interfaces (APIs) are Java, C++, and Objective-C.

Vuforia is a free software development kit for implementing mobile augmented reality. It was released by Qualcomm in 2010. With support for Android, iOS, and Unity 3D. The Vuforia platform makes it possible to write an application that can reach a large number of users on mobile devices such as smartphones, connected glasses, or tablets. Vuforia can have several functions, like text recognition.

4. Proposed approach

In this section, we present a new system based on AR composed of three modules. The first is the detection and recognition of the Arabic text of traffic signs. Second, text visualization. Third, text reading. This system helps the visually impaired or people who have difficulty reading text. This application works on mobile phones. Therefore, a smartphone camera is needed to capture an image of traffic signs. Then, store and process it. The result of this system is sound and visual output. The job is to identify textual data and display that text in a 2D image and read it aloud. To obtain this result, it is wrong to follow the process of the approach that we propose (see Fig. 1).

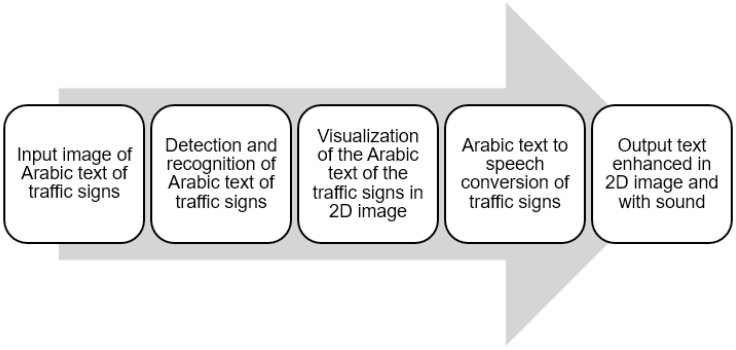

Fig. 1: Proposed architecture

These modules have been detailed in the following subsections. We use augmented reality technology to solve this text reading problem because this technology improves reality by adding digital objects. The combination of the real world and the virtual world gives us a complete and perfect environment. Therefore, if there is something in the real world that is not clear or complete enough, we correct it and accommodate it with technology to have a perfect life. This system improves text by writing it very clearly as a 2D image. Plus, you can hear it out loud. In this way, the user

can get correctly have the text. To create this system using Unity 3D software with Vuforia Engine. This application aims to give interactivity to the user when the item is selected, so that more data can be displayed and listened to.

4.1. Text detection and recognition module

In this subsection, we detail the first module of our proposed system, which is text detection and recognition using AR. The first step is for the user to hover over the image of the text they want to read using their smartphone camera.

The second step that applies to the input text image is the detection and extraction of the text box in the image. Extraction methods include rectangular box extraction, polygon extraction, and pixel-level extraction. Text detection refers to extracting the text box from an image. This study relies on the inception of a multi-scale network structure to create an AR-based Arabic text detection algorithm. The program extracts the characteristics of the Arabic text and predicts the position of the text in the image. It can better adapt to text detection and increase the efficiency of text detection. AR allows you to immediately locate the text sequence in the image. The result of text detection is shown in the following figure (Fig. 2).

Fig. 2: Image before text detection processing (a), Image after text detection processing (b)

Accurate segmentation at the pixel or character level is the major problem of classical text recognition systems based on explicit segmentation models. The subsequent recognition effect is directly influenced by the quality of the segmentation effect. Implicit segmentation, on the other hand, only requires a simple segmentation of the sample, without requiring character-level segmentation quality or precision. A word is a single string of letters. Interfering elements such as adhesion or uneven lighting between characters can prevent us from accurately segmenting each character in many natural environments. Accordingly, we apply an AR-based implicit segmentation to extract the feature sequence from the image. Traditional optical character recognition and other algorithms have matured for scanned text, but natural scene text continues to be difficult due to the influence of various factors. These factors include backgrounds, size, fonts, orientation, lighting, light text colors, textures and writing style, and viewing angles. This phase applies after the text detection phase. In this article, we rely on AR to recognize Arabic text from the natural scene. The result of text recognition is shown in the following figure (Fig. 3).

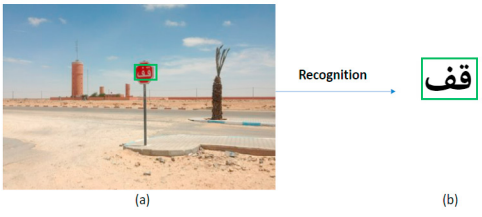

Fig. 3: Image before text recognition processing (a), Image after text recognition processing (b)

4.2. Text visualization module

For the second module of our approach, which is text visualization using AR. AR is the combination of the real world with artificial digital objects in the form of two dimensions (2D). AR is a system that combines virtual and real environments and allows it to be done interactively in real-time. Once the image is detected and recognized, additional information about the image is added to improve the result. This information is in the form of a 2D text image that will

be available to the user. Therefore, we use AR to have a more visible text display. We produce a clear visualization of the text, which is defined as a visual output. The technology used to develop this module is Unity with the help of the Vuforia engine. Unity is a cross-platform game engine. This engine is used in this system to create two-dimensional augmented reality games. Vuforia uses computer vision for image identification and tracking. Using these features allows us to add and create virtual 2D images of the real world. The virtual images appear to be part of the real-world scene. In our system, the user will get a clearer and larger text image directly on their smartphone screen. The purpose of using the Vuforia engine is to have a good visualization of the word (see Fig. 4).

Fig. 4: Image before text visualization processing (a, c), Image after text visualization processing (b, d)

4.3. Text to speech module

In this subsection, we detail the third module of our proposed system, which is a text-to-speech conversion using AR. To improve our results, we use AR to have a text-to-speech, which is defined as sound output.

Visually impaired people need a system to help them read the text and improve their quality of life. These problems can be improved by speech conversion using text-to-speech models. Many text-to-speech-based systems can help the visually impaired, but these systems have certain limitations. Therefore, a reliable solution is recommended to help such people understand the lyrics. TTS technology is designed to synthesize speech from text. Current TTS models cannot synthesize high-quality speech, and it is important to preprocess the input text to correctly map text and speech. Therefore, we tend to offer an efficient AR-based TTS system. Our speech synthesis has several advantages. For example, flexibility to modify voice characteristics, minimal footprint, and loudness. However, throughout the process of extracting and modeling speech parameters, followed by resynthesis, the naturalness of speech is significantly reduced. Our AR-based solution dramatically increases the quality of synthesized speech. End-to-end neural text-to-speech (TTS) systems have dramatically improved high-quality text-to-speech through the development of AR.

We can hear the recognized text by pressing the top-left button (Fig. 5). In the example shown in this figure, the text image is detected and recognized. Subsequently, the text will be displayed as a 2D image. Finally, the text is read clearly through this image.

5. Evaluation and experimental test

In this section, we reviewed the three test metrics to evaluate our proposed system. These three metrics are usability, efficiency, and applicability. In the rest of this section, we will give the result and the comments of the participants.

Our developed system is user-friendly and easy to use. In addition, this system is easy to use, and it does not require any training before using it. We can add that this developed system is very applicable to daily tasks, that is, it does not require conditions to be used, and it can be used at any time. This developed system is very effective for the user and especially for the visually impaired. Users can navigate and use this system anywhere. It is portable, i.e. it can be carried easily. It provides real-time speech and one can use it in your daily routine. This system can help a large community. To evaluate such a system, the best way is to use the proposed system on a group of participants. Then,

Fig. 5: Image before text to speech processing (a, c), Image after text to speech processing (b, d)

ask questions to get their feedback. Therefore, we offer our application to our participants to use after evaluation. After using it, we offer our participants (50 visually impaired) a questionnaire consisting of three questions to test their satisfaction with our system. The questionnaire is:

- **Question 1 :** How satisfied are you with the usability of this application?
- **Question 2 :** How satisfied are you with the effectiveness of this application?
- **Question 3 :** How satisfied are you with the applicability of this application?

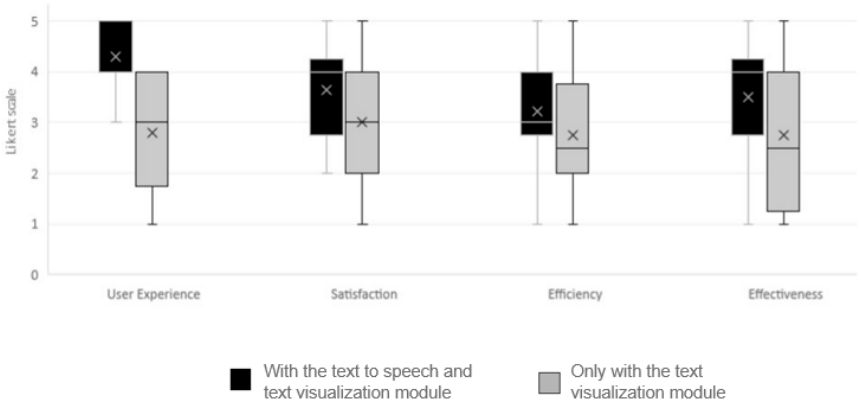

Fig. 6: Results of evaluation of our system

Figure 6 shows the results obtained after this questionnaire. As we said before, the questionnaire is based on three axes, which are usability, efficiency, and applicability. We notice that most of the participants are satisfied with this system. From this figure, we compare the difference between the experience of using our system with the text to speech and text visualization module (the black color in Fig. 6) and only with the text visualization module (the gray color in Fig. 6). Note that only the text to speech and text visualization module experience is better than only the text visualization module experience. AR technology is a good solution for the visually impaired. We conclude that the use of text-to-speech is a good complement to our system. Additionally, AR technology enhances text comprehension experiences. This technology makes it possible to superimpose layers of virtual information on a real scene to increase

the perception of reality by the user. Also, text reading is a good idea. In the context of reading text, AR has been shown to provide several benefits, namely increasing commitment to living without someone's assistance and increasing text comprehension. Especially when the text is very small, with a bad style of writing, with a very light color, or other defects.

6. Conclusion and future work

In this article, we aim to make visually impaired people as independent as possible. Current works have limitations that make them impractical and unreliable. The problem is that these methods have not yet been brought to market. Therefore, a device that can help the visually impaired is a much-desired solution. In this context, our system will enable visually impaired people to move around and help them face the world with more confidence. In this article, an AR-based mobile application is developed. It gives users the ability to read the text through an improved and better experience, mixing reality and digital content. This system consists of three axes; the detection and recognition of the text, the visualization of the text, and finally, the conversion of the text into speech. Our proposed system is developed with unity and we use the Vuforia engine. When the user uses the camera of his smartphone, the text placed in front of him will be detected, recognized, and displayed to the user very clearly and with sound. The objective of this work is, therefore, to help visually impaired people easily identify the texts in front of them. The first part of our contribution, which is the recognition of the Arabic text, is based on the Arabic text in the natural scene and more precisely on the road signs. This is an active area of research that still requires improved accuracy. The second part of our contribution is text visualization. The enhanced text visualization was important because it is the easiest and most useful for the visually impaired. The third part of our contribution is the conversion from text to speech. With this system, visually impaired people can read the text and listen to the text comfortably with less effort. This interactive audio and visual assistance system with AR is implemented in which a user can get augmented textual information through a 2D image in explanation of the displayed text and have sound. These are therefore precise and reliable solutions that can be of real help to the visually impaired. It offers an interesting way to understand their environment. In the experimental part, we observed from the results of the questionnaire that the proposed method is rather brilliant. The results obtained also increase the level of confidence and the level of comfort of the visually impaired. So they became more and more independent. In the end, the participants appreciated the result of this application. It is suggested in future works to propose a system that treats several languages at the same time. We also propose using another technology to solve these problems, such as the use of ontology [34]. We are also trying to work on a solution totally with AR such as [35] but with smart glasses.

References

[1] Abdelkarim Ben Ayed, Mohamed Ben Halima, and Adel M Alimi. Mapreduce based text detection in big data natural scene videos. In *INNS Conference on Big Data*, pages 216–223, 2015.

[2] Yahia Hamdi, Houcine Boubaker, Besma Rabhi, Wael Ouarda, and Adel Alimi. Hybrid architecture based on rnn-svm for multilingual handwriting recognition using beta-elliptic and cnn models. 2021.

[3] Mamoun Jassim Mohammed, Suphian Mohammed Tariq, and Hayder Ayad. Isolated arabic handwritten words recognition using ehd and hog methods. *Indonesian Journal of Electrical Engineering and Computer Science*, 22(2):193–200, 2021.

[4] Imran Shafi, Imtiaz Hussain, Jamil Ahmad, Pyoung Won Kim, Gyu Sang Choi, Imran Ashraf, and Sadia Din. License plate identification and recognition in a non-standard environment using neural pattern matching. *Complex & Intelligent Systems*, pages 1–13, 2021.

[5] Naseem Alrobah and Saleh Albahli. A hybrid deep model for recognizing arabic handwritten characters. *IEEE Access*, 2021.

[6] Mohammed Salemdeeb and Sarp Ertürk. Full depth cnn classifier for handwritten and license plate characters recognition. *PeerJ Computer Science*, 7:e576, 2021.

[7] Yahia Hamdi, Houcine Boubaker, and Adel M Alimi. Data augmentation using geometric, frequency, and beta modeling approaches for improving multi-lingual online handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–16, 2021.

[8] Rami Ahmed, Mandar Gogate, Ahsen Tahir, Kia Dashtipour, Bassam Al-Tamimi, Ahmad Hawalah, Mohammed A El-Affendi, and Amir Hussain. Deep neural network-based contextual recognition of arabic handwritten scripts. *Entropy*, 23(3):340, 2021.

[9] Mridul Ghosh, Himadri Mukherjee, Sk Md Obaidullah, KC Santosh, Nibaran Das, and Kaushik Roy. Lwsinet: A deep learning-based approach towards video script identification. *Multimedia Tools and Applications*, pages 1–34, 2021.

[10] Imene Ouali, Mohamed Ben Halima, and Ali Wali. Text detection and recognition using augmented reality and deep learning. In *International Conference on Advanced Information Networking and Applications*, pages 13–23. Springer, 2022.

[11] Imene Ouali, Mohamed Saifeddine Hadj Sassi, Mohamed Ben Halima, and Ali Wali. Architecture for real-time visualizing arabic words with diacritics using augmented reality for visually impaired people. In *International Conference on Advanced Information Networking and Applications*, pages 285–296. Springer, 2021.

[12] Shane Sheehan, Saturnino Luz, and Masood Masoodian. Temotopic: Temporal mosaic visualisation of topic distribution, keywords, and context. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 56–61, 2021.

[13] Mohammad Jaber Hossain and Towfik Ahmed. Augmented reality-based elementary level education for bengali character familiarization. *SN Computer Science*, 2(1):1–9, 2021.

[14] Biancamaria Mori and Carlo Gioventù. An augmented reality (ar) experience for lorenzo lotto. In *Virtual and Augmented Reality in Education, Art, and Museums*, pages 324–332. IGI Global, 2020.

[15] Imene Ouali, Mohamed Saifeddine Hadj Sassi, Mohamed Ben Halima, and WALI Ali. A new architecture based ar for detection and recognition of objects and text to enhance navigation of visually impaired people. *Procedia Computer Science*, 176:602–611, 2020.

[16] Henda Chorfi Ouertani and Lama Tatwany. Augmented reality based mobile application for real-time arabic language translation. *Communications in Science and Technology*, 4(1):30–37, 2019.

[17] Yunqiang Pei, Yadong Wu, Song Wang, Fupan Wang, Hongyu Jiang, Shijian Xu, and Jinquan Zhou. Wa vis: A web-based augmented reality text data visual analysis tool. In *2019 International Conference on Virtual Reality and Visualization (ICVRV)*, pages 11–17. IEEE, 2019.

[18] Zalfie Ardian, P Insap Santoso, and Bimo Sunarfri Hantono. Argot: Text-based detection systems in real time using augmented reality for media translator aceh-indonesia with android-based smartphones. In *Journal of Physics: Conference Series*, volume 1019, page 012074. IOP Publishing, 2018.

[19] Wei Wang, Zhikai Zhou, Yizhou Lu, Hongji Wang, Chenpeng Du, and Yanmin Qian. Towards data selection on tts data for children's speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6888–6892. IEEE, 2021.

[20] Liping Chen, Yan Deng, Xi Wang, Frank K Soong, and Lei He. Speech bert embedding for improving prosody in neural tts. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6563–6567. IEEE, 2021.

[21] Chen Zhang, Yi Ren, Xu Tan, Jinglin Liu, Kejun Zhang, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Denoispeech: Denoising text to speech with frame-level noise modeling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7063–7067. IEEE, 2021.

[22] Feng-Long Xie, Xin-Hui Li, Wen-Chao Su, Li Lu, and Frank K Soong. A new high quality trajectory tiling based hybrid tts in real time. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5704–5708. IEEE, 2021.

[23] Changfeng Gao, Gaofeng Cheng, Runyan Yang, Han Zhu, Pengyuan Zhang, and Yonghong Yan. Pre-training transformer decoder for end-to-end asr model with unpaired text data. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6543–6547. IEEE, 2021.

[24] Keisuke Matsubara, Takuma Okamoto, Ryoichi Takashima, Tetsuya Takiguchi, Tomoki Toda, Yoshinori Shiga, and Hisashi Kawai. High-intelligibility speech synthesis for dysarthric speakers with lpcnet-based tts and cyclevae-based vc. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7058–7062. IEEE, 2021.

[25] Detai Xin, Tatsuya Komatsu, Shinnosuke Takamichi, and Hiroshi Saruwatari. Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual tts. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6608–6612. IEEE, 2021.

[26] MAC Dang-Khoa, Van-Huy NGUYEN, Dinh-Nghi NGUYEN, and Kim-Anh NGUYEN. How to make text-to-speech system pronounce" voldemort": an experimental approach of foreign word phonemization in vietnamese. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6483–6487. IEEE, 2021.

[27] Song Li, Beibei Ouyang, Lin Li, and Qingyang Hong. Light-tts: Lightweight multi-speaker multi-lingual text-to-speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8383–8387. IEEE, 2021.

[28] Hwa-Yeon Kim, Jong-Hwan Kim, and Jae-Min Kim. Nn-kog2p: A novel grapheme-to-phoneme model for korean language. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7783–7787. IEEE, 2021.

[29] Cheng Gong, Longbiao Wang, Zhenhua Ling, Shaotong Guo, Ju Zhang, and Jianwu Dang. Improving naturalness and controllability of sequence-to-sequence speech synthesis by learning local prosody representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5724–5728. IEEE, 2021.

[30] YU Xinyuan and Brian Mak. Non-parallel many-to-many voice conversion by knowledge transfer from a text-to-speech model. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5924–5928. IEEE, 2021.

[31] Zilong Bai and Beibei Hu. A universal bert-based front-end model for mandarin text-to-speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6074–6078. IEEE, 2021.

[32] Junjie Pan, Lin Wu, Xiang Yin, Pengfei Wu, Chenchang Xu, and Zejun Ma. A chapter-wise understanding system for text-to-speech in chinese novels. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6069–6073. IEEE, 2021.

[33] Weiwei Jiang, Junjie Li, Minchuan Chen, Jun Ma, Shaojun Wang, and Jing Xiao. Improving neural text normalization with partial parameter generator and pointer-generator network. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7583–7587. IEEE, 2021.

[34] Imene Ouali, Faiza Ghozzi, Raouia Taktak, and Mohamed Saifeddine Hadj Sassi. Ontology alignment using stable matching. *Procedia Computer Science*, 159:746–755, 2019.

[35] Imene Ouali, Mohamed Ben Halima, and Ali Wali. Real-time application for recognition and visualization of arabic words with vowels based dl and ar. In *2022 18th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pages 678–683. IEEE, 2022.