A Quranic Dataset for Text Recognition

Idris Saleh Al-Sheikh[1], Masnizah Mohd[2]
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,
Malaysia, 43600.[12]

{[1]alshikh@gmail.com, [2]masnizah.mohd@ukm.edu.my}

Abstract : Any text recognition or Optical Character Recognition (OCR) system requires a dataset to learn how to recognize the text. Due to the lack of a standard benchmark, most of the studies in this field were conducted using private datasets without a fair comparison. In this work, we used the standard Mushaf al Madinah benchmark where there are some rules in writing style, for example, the page should start with the beginning of verse and end with the end of verse. Following these rules make the words vary in size and paragraphs on different pages. These characteristics making the recognition of the Quranic text more challenging than the normal Arabic text, where the state of the art systems fails to recognize the Quranic text. Therefore, Quranic OCR dataset is presented in this study. It contains 604 images on page level and 8927 images in text-line level. This dataset is public and free to use for the research community. The Quranic dataset would help the researchers in the field of Arabic OCR where the dataset produced in this study would be made public and free for the use of research purposes.

Keywords : *Text recognition, dataset, Quranic, Arabic OCR*

1 Introduction

Technology has become an important part of human life by making tasks easier and more efficient. One of these tasks is the Optical Character Recognition (OCR) which converts text in newspapers, magazines, and documents into digital text. Language such as Arabic requires different OCR approach due to its writing style. The Arabic language is used worldwide where it is the formal language for 25 countries and is written by more than 250 million people. The Arabic language is considered as one of the main universal sources of documents, written from right to left and uses the cursive writing style [1] as shown in Figure 1.

إِنَّا نَحۡنُ نَزَّلۡنَا ٱلذِّكۡرَ وَإِنَّا لَهُۥ لَحَٰفِظُونَ

Chapter (15) sūrat al-ḥij'r (Indeed, it is We who sent down the Qur'an and indeed, We will be its guardian)

Fig. 1. Arabic writing style

Each character may have 2 to 4 forms depending on the position of the character in the word as shown in Table 1. The presence of the hamzas (ء), dots (.) and diacritics are some of the other

INCITEST 2019, July 18, Bandung, Indonesia
Copyright © 2019 EAI
DOI 10.4108/eai.18-7-2019.2287842

features of Arabic writing which make the recognition process more complex [2]. The diacritic is a short vowel mark.

Table I. P Different Forms of the Arabic Character According to its Position in the Word

Character	Isolated	Initial	Middle	End
Alif	ا	-	-	ـا
Baa	ب	بـ	ـبـ	ـب
Kaf	ك	كـ	ـكـ	ـك
Waw	و	-	-	ـو

2 Related Works

Any text recognition or OCR system requires a dataset to learn how to recognize the text within the image, and then convert that image into digital text. Due to the lack of a standard benchmark, most of the studies in this field were conducted using private datasets without a fair comparison. Hence, although most work would showcase high accuracy results, they may not be up to scale for a large set of problems.

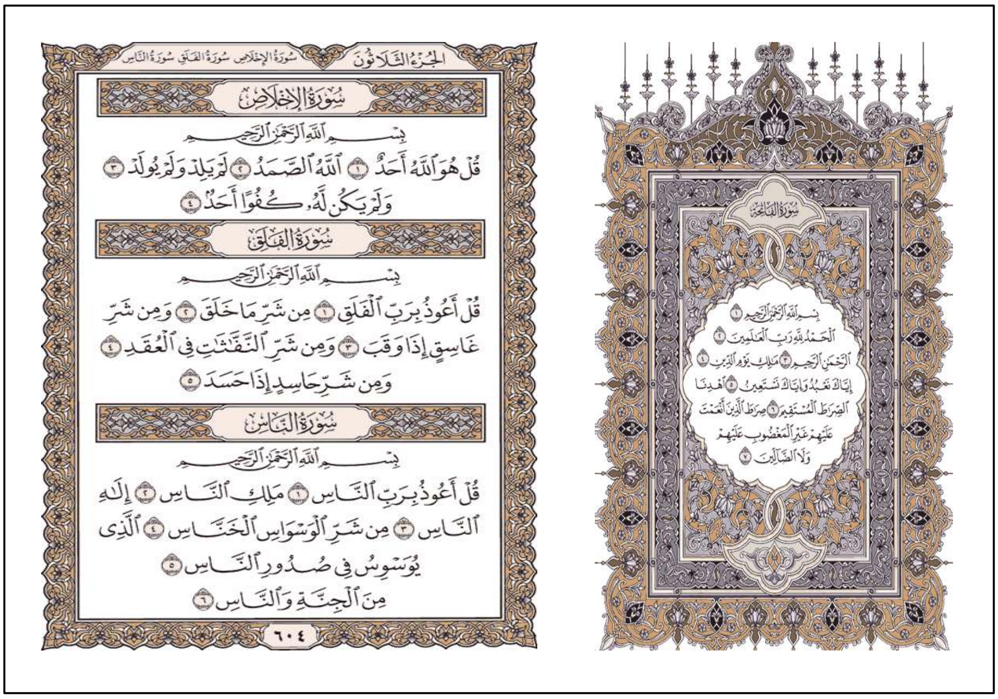
Dataset content and structure depend on the system which will use the dataset where there is dataset focusing only in digital numbers [3] and an isolated characters [4][5]. Those types of dataset usually used to train handwriting system due to the complexity of handwriting and the use of such a system is very limited. Therefore, text recognition system must recognize the whole word. However, Arabic language is a cursive language where the characters are connected together to construct a word or sub-word. Thus, sub-word dataset are been introduced [6][7] where Arabic word can be constructed from more than one group such as the word Quran قران it is one word with three-part where the first two characters are connected together as one group, the third and fourth character, each one is sub-word because there is a space between the characters. Other datasets [8][9][10] [11] used the whole word. More recent datasets especially dataset designed to train deep learning model contains image on a line level [12][13][14][15] where such a system can figure out the characters and word on the lines.

This paper introduces a public Quranic dataset on page and line level where the most unique about this dataset it focus on a diacritics text where only one dataset [11] introduce Quranic and diacritic dataset, unfortunately, this dataset is not public and it's synthetically generated on word level. In addition, what make this dataset more unique is, it based on Mushaf al Madinah benchmark which was written by the hand of Arabic calligraphy artist using the Uthmanic script.

3 Material and Methods

The first aim of this work is to build a dataset which is used in training a deep learning model. Since there is yet no known Quranic OCR dataset available to be used by the research community with its ground truth in a format that is suitable to be used in training a deep learning model. In order to achieve this goal of getting a robust Quranic OCR system, the first step would be to obtain the raw data from an authentic source. Therefore, the known benchmark of the

Mushaf Al-Madinah from the King Fahd Printing Complex was used as the source of the images which contains 604 pages as shown in Figure 2, apart from that, a Microsoft word file was also obtained from the same source which contained textual data as shown in Figure 3.



الجُزْءُ الثَّلَاثُونَ سُورَةُ الإِخْلَاصِ سُورَةُ الفَلَقِ سُورَةُ النَّاسِ

سُورَةُ الإِخْلَاصِ

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

قُلْ هُوَ ٱللَّهُ أَحَدٌ ١ ٱللَّهُ ٱلصَّمَدُ ٢ لَمْ يَلِدْ وَلَمْ يُولَدْ ٣
وَلَمْ يَكُن لَّهُۥ كُفُوًا أَحَدٌۢ ٤

سُورَةُ الفَلَقِ

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

قُلْ أَعُوذُ بِرَبِّ ٱلْفَلَقِ ١ مِن شَرِّ مَا خَلَقَ ٢ وَمِن شَرِّ
غَاسِقٍ إِذَا وَقَبَ ٣ وَمِن شَرِّ ٱلنَّفَّٰثَٰتِ فِى ٱلْعُقَدِ ٤
وَمِن شَرِّ حَاسِدٍ إِذَا حَسَدَ ٥

سُورَةُ النَّاسِ

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

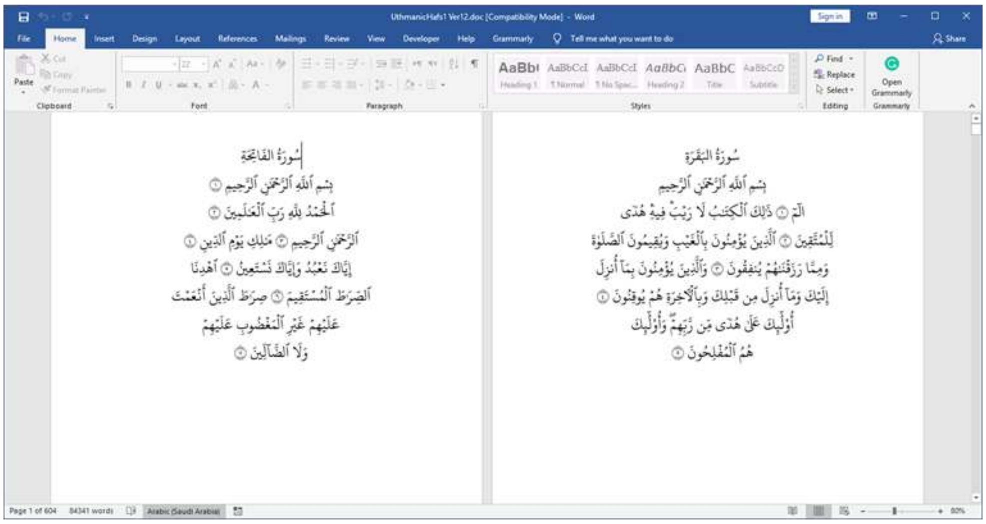
قُلْ أَعُوذُ بِرَبِّ ٱلنَّاسِ ١ مَلِكِ ٱلنَّاسِ ٢ إِلَٰهِ
ٱلنَّاسِ ٣ مِن شَرِّ ٱلْوَسْوَاسِ ٱلْخَنَّاسِ ٤ ٱلَّذِى
يُوَسْوِسُ فِى صُدُورِ ٱلنَّاسِ ٥
مِنَ ٱلْجِنَّةِ وَٱلنَّاسِ ٦

٦٠٤

سُورَةُ الفَاتِحَةِ

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ ١
ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ ٢
ٱلرَّحْمَٰنِ ٱلرَّحِيمِ ٣ مَٰلِكِ يَوْمِ ٱلدِّينِ ٤
إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ ٥ ٱهْدِنَا
ٱلصِّرَٰطَ ٱلْمُسْتَقِيمَ ٦ صِرَٰطَ ٱلَّذِينَ أَنْعَمْتَ
عَلَيْهِمْ غَيْرِ ٱلْمَغْضُوبِ عَلَيْهِمْ
وَلَا ٱلضَّآلِّينَ ٧

Fig 2. A sample image of the Mushaf Al-Madinah page



سُورَةُ الفَاتِحَةِ
بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ ١
ٱلۡحَمۡدُ لِلَّهِ رَبِّ ٱلۡعَٰلَمِينَ ٢
ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ ٣ مَٰلِكِ يَوۡمِ ٱلدِّينِ ٤
إِيَّاكَ نَعۡبُدُ وَإِيَّاكَ نَسۡتَعِينُ ٥ ٱهۡدِنَا
ٱلصِّرَٰطَ ٱلۡمُسۡتَقِيمَ ٦ صِرَٰطَ ٱلَّذِينَ أَنۡعَمۡتَ
عَلَيۡهِمۡ غَيۡرِ ٱلۡمَغۡضُوبِ عَلَيۡهِمۡ
وَلَا ٱلضَّآلِّينَ ٧

سُورَةُ البَقَرَةِ
بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ
الٓمٓ ١ ذَٰلِكَ ٱلۡكِتَٰبُ لَا رَيۡبَۛ فِيهِۛ هُدٗى
لِّلۡمُتَّقِينَ ٢ ٱلَّذِينَ يُؤۡمِنُونَ بِٱلۡغَيۡبِ وَيُقِيمُونَ ٱلصَّلَوٰةَ
وَمِمَّا رَزَقۡنَٰهُمۡ يُنفِقُونَ ٣ وَٱلَّذِينَ يُؤۡمِنُونَ بِمَآ أُنزِلَ
إِلَيۡكَ وَمَآ أُنزِلَ مِن قَبۡلِكَ وَبِٱلۡأٓخِرَةِ هُمۡ يُوقِنُونَ ٤
أُوْلَٰٓئِكَ عَلَىٰ هُدٗى مِّن رَّبِّهِمۡۖ وَأُوْلَٰٓئِكَ
هُمُ ٱلۡمُفۡلِحُونَ ٥

Fig 3. A sample of the Quranic text file

Figure 4 shows the process used to prepare the Quranic dataset. This dataset was built to train a Quranic OCR system and to study the effects of diacritics. The intention of that system is to recognize the Quranic text without the diacritics. Normally, the diacritics of the Arabic language would not be included in the written form, except for in certain situations where ambiguity may occur. In order to use this dataset for general purpose, the dataset contains 5 sets.

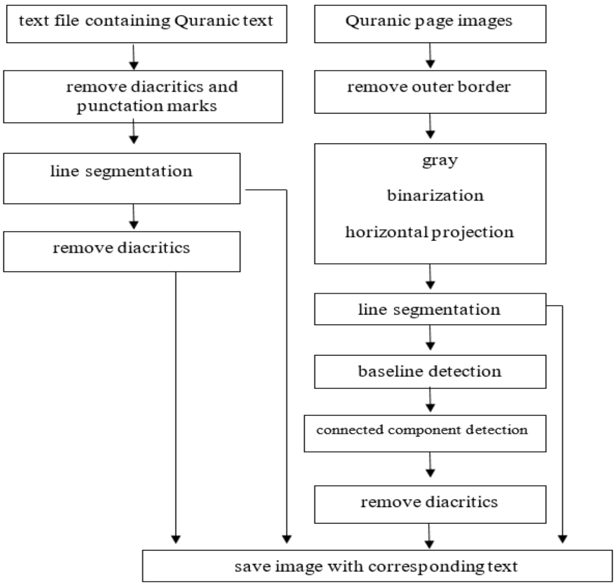


Fig 4. Methodology

The first set is on a page level where every page of the Quran are saved as an image with the corresponding text. The four remained set are saved in a line level where every line saved as an image with the ground truth text for that line. The four sets contain the same number of images. The first set contains lines images with diacritics and the ground truth text contains the equivalent text with the diacritics and verse marks as shown in Figure 5.

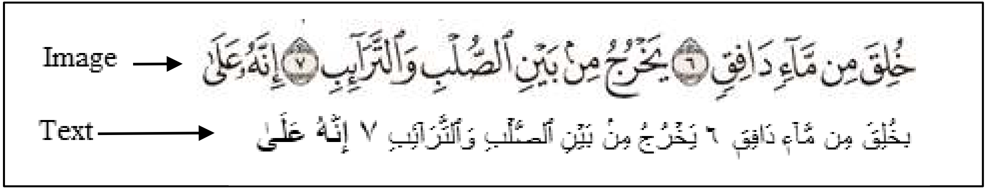


Image
Text
بخَلَقَ مِن مَّآءٍ دَافِقٍ ٦ يَخْرُجُ مِنۢ بَيْنِ ٱلصُّلْبِ وَٱلتَّرَآئِبِ ٧ إِنَّهُۥ عَلَىٰ

Fig 5. The first set

The second set contains lines images with diacritics and the ground truth text contains the equivalent text without the diacritics and verse marks as shown in Figure 6. This set can be used to recognize the text only without the diacritics.

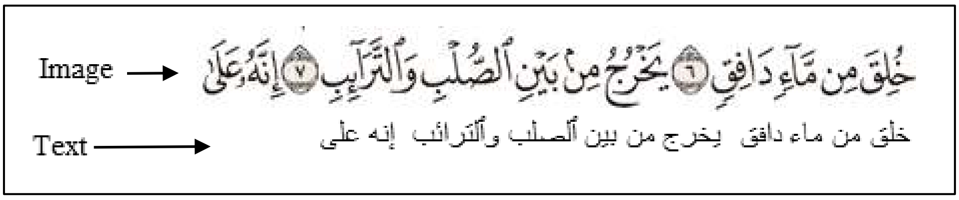


Fig 6. The second set

The third set contains lines images without diacritics and dots and the ground truth text contains the equivalent text without the diacritics, dots and verse mark as shown in Figure 7. This set used to study the effect of removing the diacritics and dots and try to minimize the number of character that the system can recognize under the assumption that removing dot and diacritics led to better accuracy, where one dot can cause an error on recognizing the characters. Arabic character can be without dot such as ر or with one dot such as ز or with two dots such as ت or three dots such as ث. The second reason is that by reducing the number of characters that the system needs to recognize will lead to fast training and small model size.

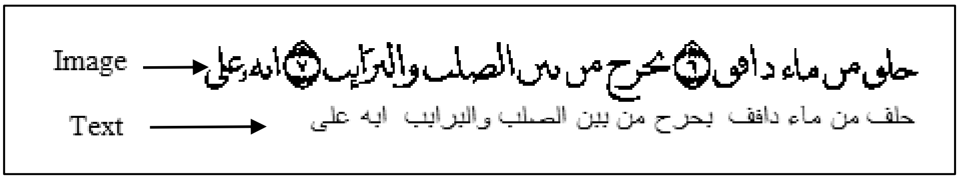


Fig 7 . The third set

The fourth set contains lines images with diacritics and the ground truth text contains the equivalent text without the diacritics, dots, and verse marks as shown in Figure 8. This set can be used as the set three, the difference is on that set the image not cleaned.

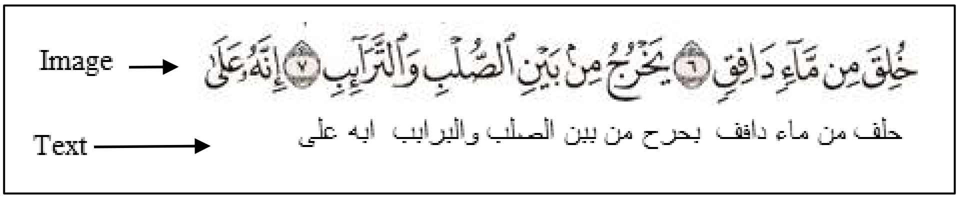


Fig 8. The fourth set

Images contain borders and therefore must be segmented in the form of separated lines before they can be used as the input for the deep learning model. Besides that, Figure 7 shows that the texts also require some normalization, where the diacritics should be removed and the verse marks. The removal of the diacritics and the verse marks, which is the rounded shape containing the verse number, must be performed since only the original texts are required. Then, every segmented image line needs to be aligned with the corresponding text line.

Mushaf Al-Madinah from the King Fahd Printing Complex is a printed copy of the Holy Quran. It contains 604 pages and the text source contains the same number of pages. The Mushaf Al-Madinah contains 84341 words, while the number of lines is 9047. In order to split the pages line by line, the horizontal projection method was used to determine the boundary of the lines.

3.1 Outer border removal

The first step is to remove the border by cropping the image where all images are of the same size as shown in Figure 9. Thus, by calculating the boundary of the border for one image, the method was used in cropping the rest of the image borders, with the exception of the first and second pages where the borders are different and therefore required for the cropping to be performed manually.

B. Parameter of the simulation.

Three switching control signals, i.e. PWL, PWM or SPWM can be applied. This section analyzes the THD and output voltage amplitude when using one of third switching signals. Fig. 8 shows the time domain output voltage simulation results using third switching signals.

Table I. Parameters Simulation Of 3 Level Cascaded Multilevel Inverter

Trafoless/Trafo	Trafo		Trafoless
input Voltage	48V		220V
Output Voltage	61V		214V
Frequency	50Hz		50Hz
Output Power	27Watt		229Watt
Var./Parameter	**LCL 2Leg's**	**Trafo**	**LCL 2Leg**
L1	8 mH	-	8 mH

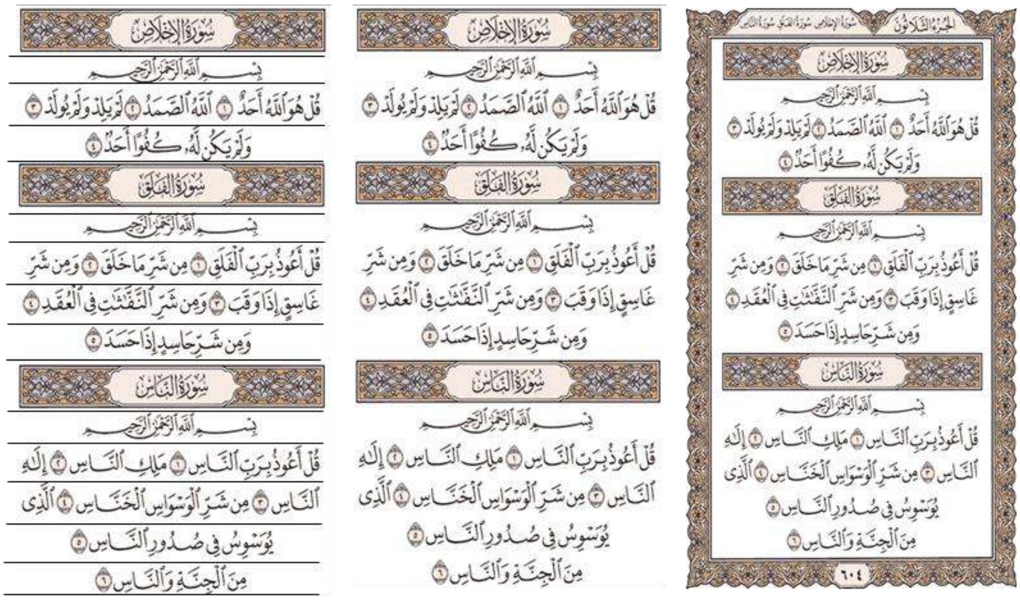


Fig 9. Line Segmentations

3.2 Line Segmentations

The second step is to segment the page image to lines images by calculating the horizontal projection profile for the image, which means that the image must be converted into white and black colours before the calculation of the sum of the black pixel for every row of the image could be conducted. Figure 10 shows the horizontal projection for the text line image, and in doing that the colour image should be converted into grayscale. Next, a threshold is applied in converting every pixel of the image into white colour. This is represented by the pixel value of 255 for white colour or a pixel value of 0 for black colour by converting the image into a grey level. This means to convert the three-color channels into one channel with pixel value between 0 to 255. To convert the grey image into binary, the threshold should be used. If the value of every pixel in the image is greater than the threshold, then the pixel value will be 255. However, if the value is less than the threshold, then the pixel value will be 0. This is followed with the calculation of the sum of the black pixel for each row of the image, as shown in Figure 10 that the baseline for every text line could be determined from the horizontal projection. The maximum value represents the baseline while the minimum value represents the end of the line where the text line can be splatted from.

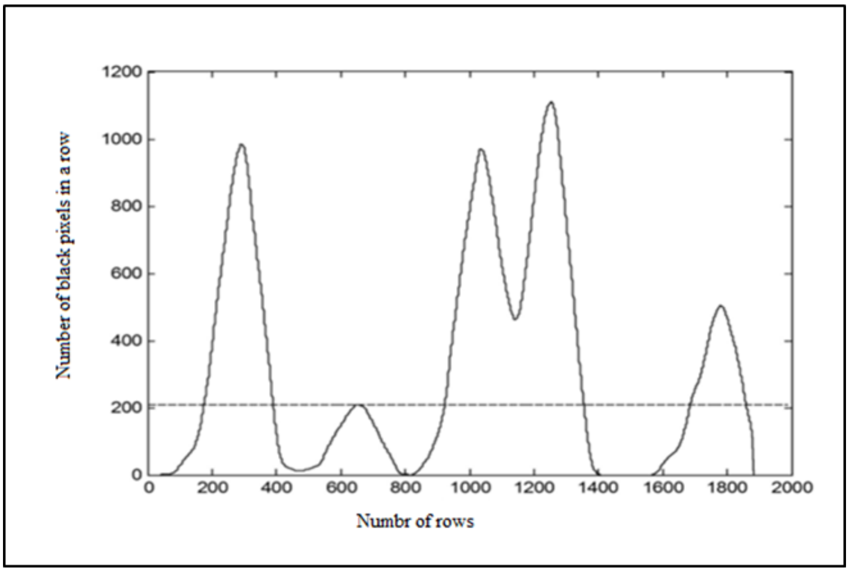
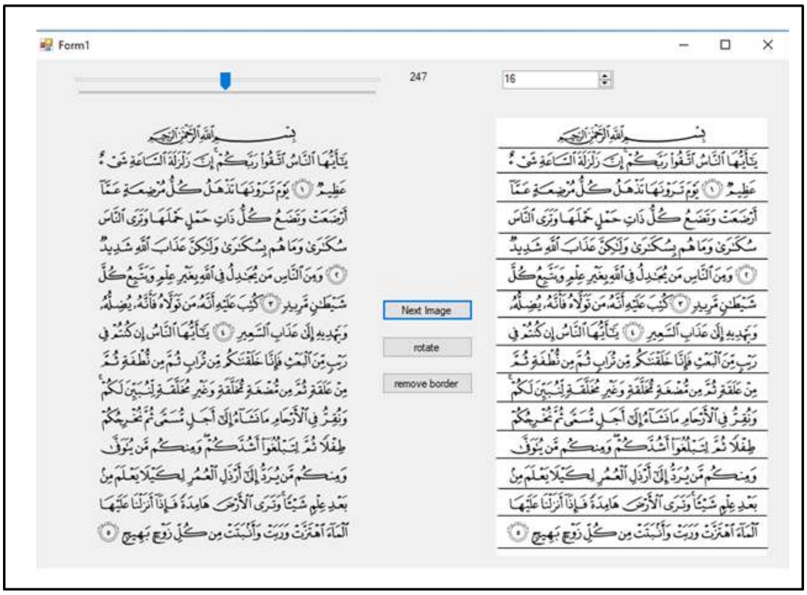


Fig 10. Line Segmentations



بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ
يَٰٓأَيُّهَا ٱلنَّاسُ ٱتَّقُواْ رَبَّكُمۡۚ إِنَّ زَلۡزَلَةَ ٱلسَّاعَةِ شَيۡءٌ
عَظِيمٞ ١ يَوۡمَ تَرَوۡنَهَا تَذۡهَلُ كُلُّ مُرۡضِعَةٍ عَمَّآ
أَرۡضَعَتۡ وَتَضَعُ كُلُّ ذَاتِ حَمۡلٍ حَمۡلَهَا وَتَرَى ٱلنَّاسَ
سُكَٰرَىٰ وَمَا هُم بِسُكَٰرَىٰ وَلَٰكِنَّ عَذَابَ ٱللَّهِ شَدِيدٞ
٢ وَمِنَ ٱلنَّاسِ مَن يُجَٰدِلُ فِي ٱللَّهِ بِغَيۡرِ عِلۡمٖ وَيَتَّبِعُ كُلَّ
شَيۡطَٰنٖ مَّرِيدٖ ٣ كُتِبَ عَلَيۡهِ أَنَّهُۥ مَن تَوَلَّاهُ فَأَنَّهُۥ يُضِلُّهُۥ
وَيَهۡدِيهِ إِلَىٰ عَذَابِ ٱلسَّعِيرِ ٤ يَٰٓأَيُّهَا ٱلنَّاسُ إِن كُنتُمۡ فِي
رَيۡبٖ مِّنَ ٱلۡبَعۡثِ فَإِنَّا خَلَقۡنَٰكُم مِّن تُرَابٖ ثُمَّ مِن نُّطۡفَةٖ ثُمَّ
مِنۡ عَلَقَةٖ ثُمَّ مِن مُّضۡغَةٖ مُّخَلَّقَةٖ وَغَيۡرِ مُخَلَّقَةٖ لِّنُبَيِّنَ لَكُمۡۚ
وَنُقِرُّ فِي ٱلۡأَرۡحَامِ مَا نَشَآءُ إِلَىٰٓ أَجَلٖ مُّسَمّٗى ثُمَّ نُخۡرِجُكُمۡ
طِفۡلٗا ثُمَّ لِتَبۡلُغُوٓاْ أَشُدَّكُمۡۖ وَمِنكُم مَّن يُتَوَفَّىٰ
وَمِنكُم مَّن يُرَدُّ إِلَىٰٓ أَرۡذَلِ ٱلۡعُمُرِ لِكَيۡلَا يَعۡلَمَ مِنۢ
بَعۡدِ عِلۡمٖ شَيۡـٔٗاۚ وَتَرَى ٱلۡأَرۡضَ هَامِدَةٗ فَإِذَآ أَنزَلۡنَا عَلَيۡهَا
ٱلۡمَآءَ ٱهۡتَزَّتۡ وَرَبَتۡ وَأَنۢبَتَتۡ مِن كُلِّ زَوۡجِۭ بَهِيجٖ ٥

بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ
يَٰٓأَيُّهَا ٱلنَّاسُ ٱتَّقُواْ رَبَّكُمۡۚ إِنَّ زَلۡزَلَةَ ٱلسَّاعَةِ شَيۡءٌ
عَظِيمٞ ١ يَوۡمَ تَرَوۡنَهَا تَذۡهَلُ كُلُّ مُرۡضِعَةٍ عَمَّآ
أَرۡضَعَتۡ وَتَضَعُ كُلُّ ذَاتِ حَمۡلٍ حَمۡلَهَا وَتَرَى ٱلنَّاسَ
سُكَٰرَىٰ وَمَا هُم بِسُكَٰرَىٰ وَلَٰكِنَّ عَذَابَ ٱللَّهِ شَدِيدٞ
٢ وَمِنَ ٱلنَّاسِ مَن يُجَٰدِلُ فِي ٱللَّهِ بِغَيۡرِ عِلۡمٖ وَيَتَّبِعُ كُلَّ
شَيۡطَٰنٖ مَّرِيدٖ ٣ كُتِبَ عَلَيۡهِ أَنَّهُۥ مَن تَوَلَّاهُ فَأَنَّهُۥ يُضِلُّهُۥ
وَيَهۡدِيهِ إِلَىٰ عَذَابِ ٱلسَّعِيرِ ٤ يَٰٓأَيُّهَا ٱلنَّاسُ إِن كُنتُمۡ فِي
رَيۡبٖ مِّنَ ٱلۡبَعۡثِ فَإِنَّا خَلَقۡنَٰكُم مِّن تُرَابٖ ثُمَّ مِن نُّطۡفَةٖ ثُمَّ
مِنۡ عَلَقَةٖ ثُمَّ مِن مُّضۡغَةٖ مُّخَلَّقَةٖ وَغَيۡرِ مُخَلَّقَةٖ لِّنُبَيِّنَ لَكُمۡۚ
وَنُقِرُّ فِي ٱلۡأَرۡحَامِ مَا نَشَآءُ إِلَىٰٓ أَجَلٖ مُّسَمّٗى ثُمَّ نُخۡرِجُكُمۡ
طِفۡلٗا ثُمَّ لِتَبۡلُغُوٓاْ أَشُدَّكُمۡۖ وَمِنكُم مَّن يُتَوَفَّىٰ
وَمِنكُم مَّن يُرَدُّ إِلَىٰٓ أَرۡذَلِ ٱلۡعُمُرِ لِكَيۡلَا يَعۡلَمَ مِنۢ
بَعۡدِ عِلۡمٖ شَيۡـٔٗاۚ وَتَرَى ٱلۡأَرۡضَ هَامِدَةٗ فَإِذَآ أَنزَلۡنَا عَلَيۡهَا
ٱلۡمَآءَ ٱهۡتَزَّتۡ وَرَبَتۡ وَأَنۢبَتَتۡ مِن كُلِّ زَوۡجِۭ بَهِيجٖ ٥

Fig 11. Screenshot of image segmentation into lines

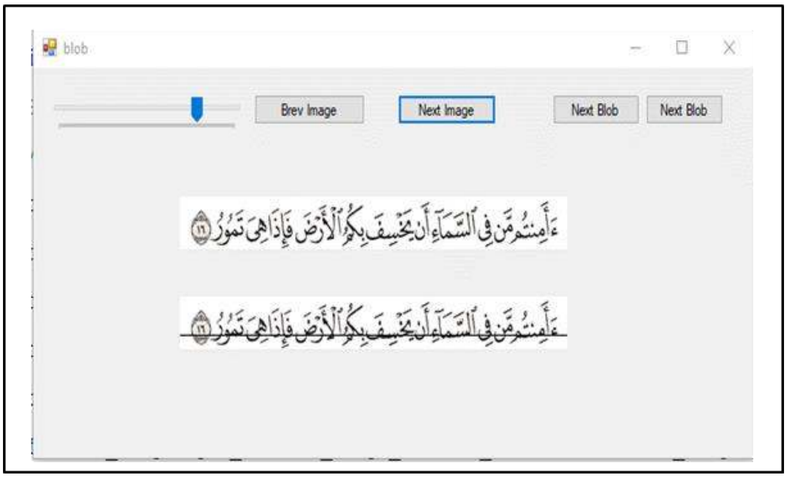
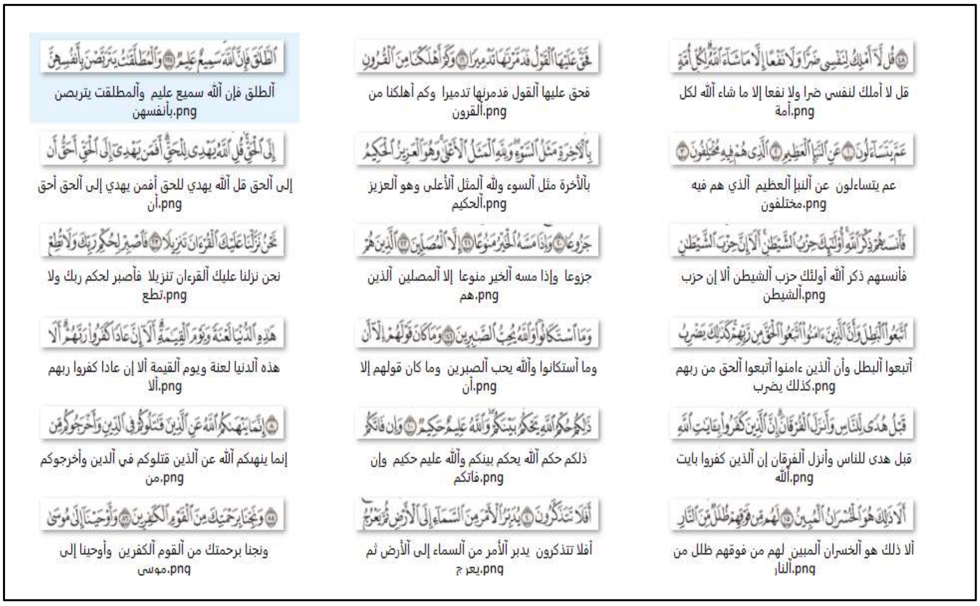


Fig 12. Baseline

When a page is segmented into lines (Figure 11), the page number and the line number are added to the name of the split line so that the ground truth text could be mapped (Figure 12). This is achieved by tracking the page and line numbers and then applying the text normalization and saving the ground truth text as the image name as shown in Figure 13.



ٱلطلق فإن ٱلله سميع عليم وٱلمطلقت يتربصن بأنفسهن.png

فحق عليها ٱلقول فدمرنها تدميرا وكم أهلكنا من ٱلقرون.png

قل لا أملك لنفسي ضرا ولا نفعا إلا ما شاء ٱلله لكل أمة.png

إلى ٱلحق قل ٱلله يهدي للحق أفمن يهدي إلى ٱلحق أحق أن.png

بٱلأخرة مثل ٱلسوء ولله ٱلمثل ٱلأعلى وهو ٱلعزيز ٱلحكيم.png

عم يتساءلون عن ٱلنبإ ٱلعظيم ٱلذي هم فيه مختلفون.png

نحن نزلنا عليك ٱلقرءان تنزيلا فٱصبر لحكم ربك ولا تطع.png

جزوعا وإذا مسه ٱلخير منوعا إلا ٱلمصلين ٱلذين هم.png

فأنسىهم ذكر ٱلله أولئك حزب ٱلشيطن ألا إن حزب ٱلشيطن.png

هذه ٱلدنيا لعنة ويوم ٱلقيمة ألا إن عادا كفروا ربهم ألا.png

وما ٱستكانوا وٱلله يحب ٱلصبرين وما كان قولهم إلا أن.png

ٱتبعوا ٱلبطل وأن ٱلذين ءامنوا ٱتبعوا ٱلحق من ربهم كذلك يضرب.png

إنما ينهىكم ٱلله عن ٱلذين قتلوكم في ٱلدين وأخرجوكم من.png

ذلكم حكم ٱلله يحكم بينكم وٱلله عليم حكيم وإن فاتكم.png

قبل هدى للناس وأنزل ٱلفرقان إن ٱلذين كفروا بايت ٱلله.png

ونجنا برحمتك من ٱلقوم ٱلكفرين وأوحينا إلى موسى.png

أفلا تتذكرون يدبر ٱلأمر من ٱلسماء إلى ٱلأرض ثم يعر ج.png

ألا ذلك هو ٱلخسران ٱلمبين لهم من فوقهم ظلل من ٱلنار.png

Fig 13. Screenshot of dataset images with ground truth text

3.3 Diacritics removal

To remove the diacritics and dots from the images, morphological process on images is used to detect the connected component in the image. Besides that, the baseline is used to draw lines over the images so that every part that is not connected to the baseline will be removed as shown in Figure 14. Some diacritics have not been removed because those diacritics are connected with the words, and thus the deep learning model will consider these diacritics as noise since they are not present in the ground truth text. Furthermore, Figure 14 also shows some characters which are positioned above the baseline, for example, the character (أ) in the fourth word (السماء). To solve this issue, a threshold is added to the baseline thickness. As a result of this, as shown in Figure 15, the missing character now is present after the threshold is added to the baseline thickness.

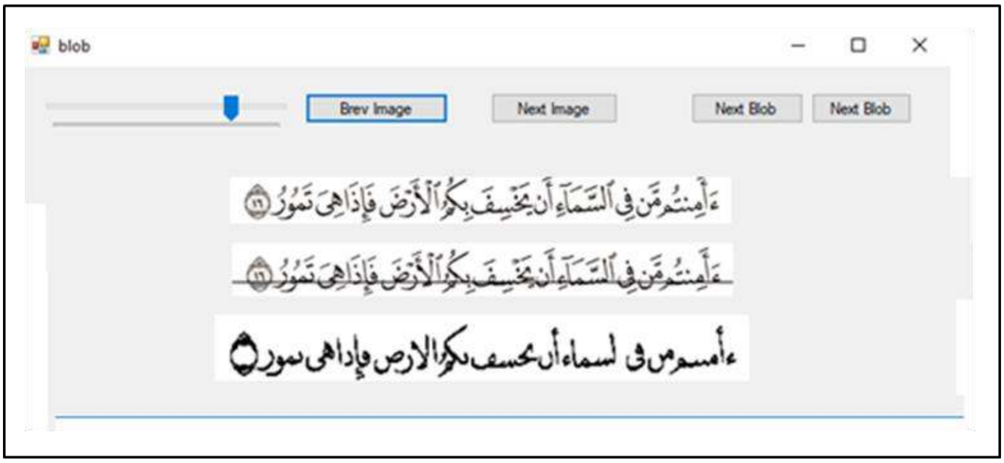


Fig 14. Screenshot of the removal of diacritics and dots

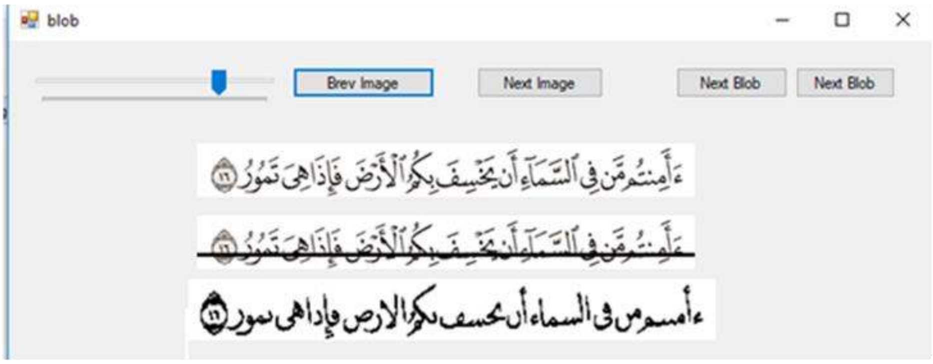


Fig 15. Screenshot of the removal of diacritics and dots with threshold baseline

As shown in Figure 15, the output image is clear from diacritics and dots. The ground truth text must be equal to the text on the image, therefore the text needs to be normalized. The normalization process for the text is performed by changing some characters based on some rules. As mentioned previously, the Arabic characters can be in the form of four states depending on the position of the characters in a word, either isolated, in the beginning, in the middle or at the end of the word. After the dots are removed, more than one character will look the same or have small differences. Hence, some characters have been compounded together based on the shape of the characters in the word as shown in Table 2. By applying these rules, the normalized text is therefore provided and this text can now be used as the ground truth to the clean image.

Table I. Normal and normalized character based on character shapes

	Normal Form				**New Form**			
Normalized Character	Isolated	Initial	Middle	End	Isolated	Initial	Middle	End
	ا	ا	ﺎ	ﺎ				
	آ	آ	–	ﺄ				
ا	أ	أ	ﺄ	ﺄ	ا	ا	ﺎ	ﺎ
	إ	إ	–	–				
	ٱ	ٱ	–	–				
	ب	ﺑ	ﺒ	ﺐ				
	ت	ﺗ	ﺘ	ﺖ	ب			
ب	ث	ﺛ	ﺜ	ﺚ		ﺑ	ﺒ	
	ن	ﻧ	ﻨ	ﻦ	ن			ﺐ
	ي	ﻳ	ﻴ	ﻲ	ي			ﺐ
	ج	ﺟ	ﺠ	ﺟ				
ح	ح	ﺣ	ﺤ	ﺣ	ح	ﺣ	ﺤ	ﺣ
	خ	ﺧ	ﺨ	ﺧ				
د	د	د	ﺪ	ﺪ	د	ﺪ	ﺪ	ﺪ
	ذ	ذ	ﺬ	ﺬ				
ر	ر	ر	ﺮ	ﺮ	ر	ر	ﺮ	ﺮ
	ز	ز	ﺰ	ﺰ				
س	س	ﺳ	ﺴ	ﺲ	س	ﺳ	ﺴ	ﺲ
	ش	ﺷ	ﺸ	ﺶ				
ص	ص	ﺺ	ﺼ	ﺺ	ص	ﺺ	ﺼ	ﺺ
	ض	ﺾ	ﻀ	ﺾ				
ط	ط	ﻂ	ﻂ	ﻂ	ط	ط	ﻂ	ﻂ
	ظ	ﻆ	ﻆ	ﻆ				
ع	ع	ﻋ	ﻌ	ﻊ	ع	ﻊ	ﻌ	ﻊ
	غ	ﻏ	ﻐ	ﻎ				

ف	ف ق	ﻓ ﻗ	ﻔ ﻘ	ﻒ ﻖ	ف	ﻒ	ﻔ	ﻒ
ك ل	ك ل	ﻛ ﻟ	ﻜ ﻠ	ﻚ ﻞ	ل	ﻛ ﻟ	ﻜ ﻠ	ﻞ
م	م	ﻣ	ﻤ	ﻢ	م	ﻣ	ﻤ	ﻢ
ه ة	ه ة	ﻫ –	ﻬ –	ﻪ ﺔ	ه	ﻫ	ﻬ	ﻪ
و	و	ﻮ	ﻮ	ﻮ	و	ﻮ	ﻮ	ﻮ

4. Conclusion

The Quranic OCR dataset is presented in this study, where the Holy Quran is considered as the most important Islamic and Arabic book. It was revealed and written in the Arabic language, unfortunately as the previous research report poor accuracy in recognizing the Quranic text. Experiment to measure the accuracy of two of the state of art OCR system ABBYY FineReader and Tesseract, the two systems were had been unable to recognize the Quranic text, while a system trained on the proposed dataset obtained an accuracy of 98% on validation data and obtained Word Error Rate (WER) 2.66 and Character Error Rate (CER) 0.72 in the test dataset. The Quranic dataset would help the researchers in the field of Arabic OCR where the dataset produced in this study would be made public and free for the use of research purposes.

References

[1] K. Addakiri and M. Bahaj,: “On-line handwritten arabic character recognition using artificial neural network,” Int. J. Comput. Appl., vol. 55, no. 13, 2012.

[2] O. Al-Jarrah, :“A New Algorithm for Arabic Optical Character Recognition,” Proc. 5th WSEAS Int Conf Signal Process. Robot. Autom., vol. 2006, no. April, pp. 211–224, 2006.

[3] S. M. Awaidah and S. A. Mahmoud,: “A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models,” Signal Processing, vol. 89, no. 6, pp. 1176–1184, 2009.

[4] A. Asiri and M. S. Khorsheed,: “Automatic Processing of Handwritten Arabic Forms using Neural Networks.,” in IEC (Prague), pp. 313–317. 2005

[5] A. Lawgali, M. Angelova, and A. Bouridane,: “HACDB: Handwritten Arabic characters database for automatic character recognition,” in Visual Information Processing (EUVIP), 2013 4th European Workshop on, pp. 255–259. 2013

[6] Y. Al-Ohali, M. Cheriet, and C. Suen, :“Databases for recognition of handwritten Arabic cheques,” Pattern Recognit., vol. 36, no. 1, pp. 111–121, 2003.

[7] B. Bataineh,: “A Printed PAW Image Database of Arabic Language for Document Analysis and Recognition,” J. ICT Res. Appl., vol. 11, no. 2, pp. 200–212, 2017.

[8] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, and H. Amiri,: “IFN/ENIT-database of handwritten Arabic words,” in Proc. of CIFED, vol. 2, pp. 127–136. 2002

[9] S. Al-Ma’adeed, D. Elliman, and C. A. Higgins,: “A data base for Arabic handwritten text recognition research,” in Frontiers in Handwriting Recognition, Proceedings. Eighth International Workshop on, pp. 485–489. 2002

[10] O. Zayene, S. M. Touj, J. Hennebert, R. Ingold, and N. E. Ben Amara, :“Open Datasets and Tools for Arabic Text Detection and Recognition in News Video Frames,” J. Imaging, vol. 4, no. 2, p. 32, 2018.

[11] M. Badry, H. Hassan, H. Bayomi, and H. Oakasha, :“QTID: Quran Text Image Dataset,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 3, pp. 385–391, 2018.

[12] N. Sabbour and F. Shafait, :“A segmentation-free approach to Arabic and Urdu OCR,” in Document Recognition and Retrieval XX , vol. 8658, p. 86580N. 2013

[13] S. A. Mahmoud et al.,: “KHATT: An open Arabic offline handwritten text database,” Pattern Recognit., vol. 47, no. 3, pp. 1096–1112, 2014.

[14] S. Yousfi, S.-A. Berrani, and C. Garcia,: “ALIF: A dataset for Arabic embedded text recognition in TV broadcast,” in Document Analysis and Recognition (ICDAR), 13th International Conference on, pp. 1221–1225. , 2015

[15] O. Zayene, J. Hennebert, S. M. Touj, R. Ingold, and N. E. Ben Amara,: “A dataset for Arabic text detection, tracking and recognition in news videos-AcTiV,” in Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, pp. 996–1000. 2015