Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Multi Head Attention Transformer for Arabic Scene Images Text Recognition

Oualid KHIAL[1], Fatma BOUFERRA[2]
[1] *LISYS Laboratory, Exact Sciences Faculty, Mustapha STAMBOULI University, Cheikh El Khaldi Street, Mascara, 29000, Mascara, Algeria*

ARTICLE INFO

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

The worldwide video library continues to expand rapidly, which creates an increasing need for modern and reliable techniques for video processing and text indexing. In this paper, we introduce a new implementation of the Transformer architecture for scene text recognition. This work comes from a comparative study between two approaches: using convolutional feature maps as input to the Transformer encoder, and fully removing any CNN component. During training, we used almost all available public datasets; however, they were still not enough because of the significant lack of large-scale and diverse datasets for this task. This challenge led us to create and publish a new artificial dataset called **IYaD**. The IYaD dataset currently contains around 1,400,000 images for one font and the same scale for 16 additional fonts. Each image is provided in three different versions and includes Arabic labels, Latin transcription, and the text content. The experimental results show that our Transformer-based ASTR model surpasses state-of-the-art methods, especially when trained on the IYaD dataset, establishing new benchmarks in accuracy and robustness. We believe that this dataset demonstrates the importance and potential of artificially created datasets, and it may encourage similar dataset generation in other research domains.

Keywords: Transformer; Arabic Data set; text recognition; AI; Neural networks; deep learning; machine learning

1. INTRODUCTION

Treating and indexing videos require using all the available information, whether it is present in the video itself or found as metadata associated to it. This means that any piece of information—no matter how small—can contribute to a richer and more meaningful understanding of the video's content. Among the most valuable of these are the embedded text frames, which often carry important contextual clues, such as names, locations, dates, product details, or instructions that help in identifying and describing what is being depicted. By accurately extracting and interpreting this textual information, the indexing process becomes more precise, which in turn improves searchability, accessibility, and the overall usefulness of the video content.

On the other hand, the growing number of Arabic speakers worldwide—which now comprises nearly one-quarter of the global population—has increased the demand for efficient tools and systems that can handle Arabic language processing in real-world applications. However, dealing with Arabic text introduces several layers of complexity. The Arabic script is written from right to left and is characterized by a large set of ligatures, diacritical marks, and a wide variety of fonts. In addition, each Arabic character can appear in multiple forms depending on its position within the word (Table 1), and several characters have similar shapes, making them visually hard to distinguish (Table 2). These linguistic and structural characteristics alone pose significant challenges.

Isolated	connected	double connected	transformed
ج	جـ	ـجـ	None
ت	تـ	ـتـ	ـة

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Isolated	connected	double connected	transformed
ك	كـ	كـ	None

Table 1 Single char shape variations

Moreover, when this complexity is combined with the nature of real-world scenes such as videos with noisy or cluttered backgrounds, varying lighting conditions, distortions, and low resolution (see Figure 1) makes the task of recognizing Arabic text becomes even more difficult. The absence of well-structured, standardized, and reliable datasets further complicates the development of robust models for Arabic Scene Text Recognition (ASTR). As a result, extracting and interpreting Arabic text from video frames is not just a simple technical step; it is a demanding and intricate process that requires careful consideration of linguistic details, image quality issues, and the lack of sufficient training resources.

Figure 1 Noisy real world images examples

From another perspective, A quick check on the world wide publicly available data sets, will highlight the fact that there is an extreme lack of reliable data set for ASTR regardless of the large number of Arabic speakers around the world and the continuous growth of video data from Arabic sources.

With all this facts in mind, This paper propose a new Transformer based architecture dedicated to approaching the ASTR problem from a new perspective abandoning the use of the classic Recurrent Neural Networks (RNN) and the use of Convolutional neural networks (CNN) exploiting the Transformer ability to find and recognize pasterns in long sequences of data.

In this work, we utilize the **IYad** dataset (Oualid et al., n.d.), which we have previously developed and made publicly available to support research on Arabic text recognition. At the time of its release, the dataset comprised approximately 2.8 million images, with additional samples being continuously generated. Each image contains either a single Arabic word (Iyad-W) or a full meaningful sentence (Iyad-S). To enhance variability and robustness, every image is provided in three different formats: white background, randomly colored background, and real-world background. Moreover, the dataset includes identical samples rendered in sixteen distinct Arabic fonts, resulting in a total of 44.8 million images. This scale and diversity make it particularly valuable for data augmentation and the training of deep learning models.

In the next sections, we present and analyses related works in term of methods or datasets in Section 2, We present the data set with more details in Section 5. All input and execution scenarios are detailed in section 3.

Isolated	beginning	middle	English sound
ب	بـ	ـبـ	Ba
ت	تـ	ـتـ	Ta
ي	يـ	ـيـ	Ya

Table 2 Characters similarity

2. RELATED WORKS

Text recognition is not a new field of research, Actually many methods and techniques and has been proposed in the last decades; As far as we know, they can be classified in two categories: classic methods which aims to manually extract the text features by asking the question how do humans read a text in an image than write algorithms to automate the feature extraction and to more modern approach relying on artificial neural networks and deep learning techniques.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In the context of classics methods, (Wang et al., 2011) construct and evaluate two systems. The first is a two-stage pipeline consisting of text detection followed by a leading OCR engine. The second is a system rooted in generic object recognition. (Anthimopoulos et al., 2007) propose a method based an edge candy edge detector and connected components.
The second category is machine leaning based approaches where algorithms like KNN, SVM, K-means clustering are used usually after a prepossessing step.

An orientation free method was proposed by (Shivakumara et al., 2011).They use a Fourier-Laplacian filter and K-means clustering to identify candidate text regions. split the ASTR pipe line into detection step and recognition, in the first step, they use maximally stable extremal regions and color clustering as connected components extractor than filter text areas using visual saliency and other prior information. For the recognition step, they use SVM based framework and a character classifier.

An Hybrid work focused on the identification of words of situation. (Panhwar et al., 2019) use classic edge detection techniques to detect the signboard edges than extract the word inside. lately an Artificial Neural Network is used for the classification and recognition of the text extracted.

An other related work focused on English text has been proposed by (Liao et al., 2019), they approach the problem from a two-dimensional perspective using a Character Attention Fully Convolutional Network (CA-FCN) which use attention mechanism on the character level.

Toward a more modern and deep learning based approaches, propose a to use a new type of recurrent neural networks (RNN) which proves to outperform a state-of-the-art of handwritten text recognition in the time the paper was submitted (Graves et al., 2009).

More recent approaches combined Multi-Dimensional Long Short Term Memory (MDLSTM) which is a more sophisticated class of RNN with Connectionist Temporal Classification (CTC) to do a more effective sequence labeling and demonstrated a very promising result (Yousfi et al., 2015b) (Zayene et al., 2017).

A highly related work based on transformer was introduced by , they were the first to spot the transformers problem when dealing with text images. they spot the problem of the model estimated alignments corruption and they call it attention drift. they develop a mechanism to force the attention network on the region in interest and get promising results (Cheng et al., 2017).

In the context of datasets, The earliest Arabic text datasets primarily focused on handwritten documents, covering both historical and modern Arabic manuscripts.

One of the early presented and Arabic town and village names handwritten dataset. The text written is from 937 Tunisian town/village names. A pre-label assigned to each file consists of the postcode in a sequence of Numeric Character References, which stored in the UPX file format. An XML file including trajectory information and a plot image of the word trajectory are also generated. Additional information about the writer can also be provided (El Abed et al., 2011).

(Mezghani et al., 2012)presented an Arabic Handwritten Text Images Database written by Multiple Writers (AHTID/MW). The AHTID/MW contains 3710 text lines and 22896 words written by 53 native writers of Arabic along with the ground truth annotation for each image.

(Al-Ohali et al., 2003) dedicated an effort towards the development of Arabic cheque databases for research in the recognition of hand-written Arabic cheques. they provided a solid validation procedure including grammars and algorithms used to verify the correctness of the tagging process.

Dealing with printed text, two datasets have been developed to assess Arabic text recognition in various contexts. The Arabic Printed Text Image database (APTI) (Slimane et al., 2009) contains 45,313,600 Arabic printed word images with 10 different fonts, sizes, and styles, such as italic and bold. APTI is particularly useful for evaluating Arabic text recognition in screen captures or images extracted from PDF documents, but it only consists of synthetic text images with a clean white background.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In the same context, the ATID/MF dataset (Arabic Printed Text Image Database/Multi-Font) (Jaiem et al., 2013) is based on 387 gray scale scanned pages of Arabic printed documents from a Tunisian newspaper's official website. It contains 27,402 samples and ground truth labels. The database is freely available to interested researchers. The PATD database presented by (RAMDAN et al., 2013) contains eight hundred and ten images scanned in grayscale format and different resolutions, leading to two thousand and nine hundred and fifty-four images under varying capture conditions (blurred, at different angles and in different light. Toward more recent contributions, The ALiF (Yousfi et al., 2015a) data set is composed of a large number of manually annotated text images that were extracted from Arabic TV broadcast. the data set is well structured but the number of images is not enough and it is not being updated as far as we know. The most recent as far as we know is the AcTiV 2.0 dataset, it is specifically designed for the development and assessment of Arabic video text detection and recognition systems. AcTiV 2.0 comprises 189 video clips, forming the basis for generating 4063 key frames for detection and 10,415 cropped text images for recognition. Additionally, AcTiV 2.0 comes with open-source annotation and evaluation tools, facilitating standardization and validation efforts (Zayene et al., 2018).

3. PROPOSED APPROACH

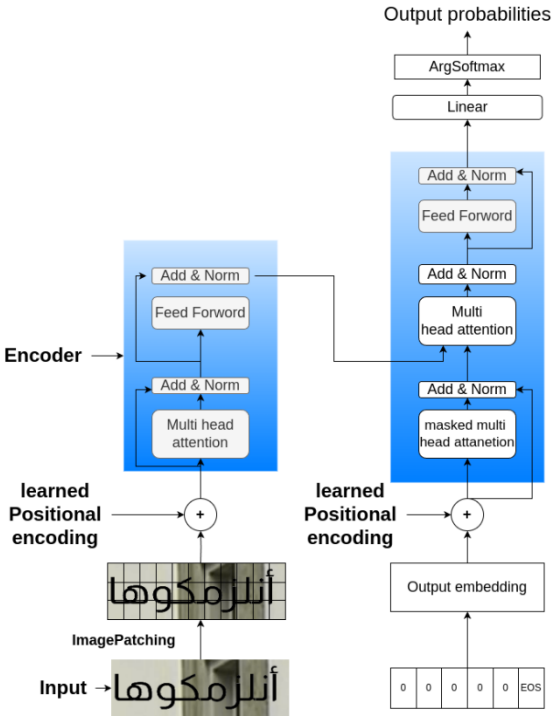

Figure 2 Proposed model

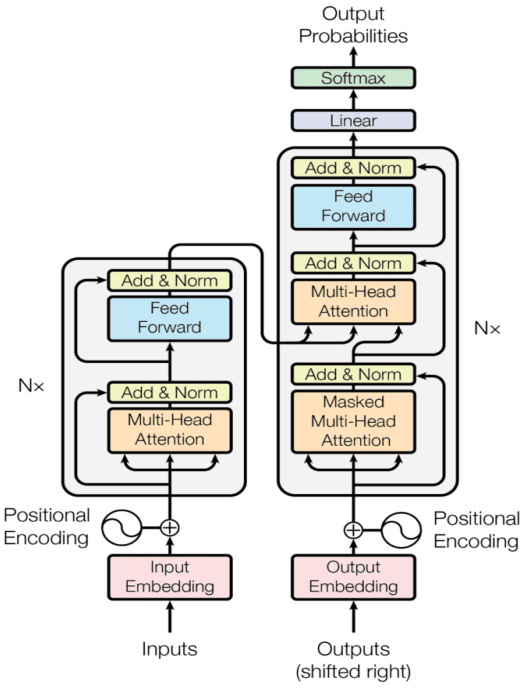

Figure 3 Original transformer model

Arabic scene text recognition and almost any text or character recognition has always been considered as a computer vision problem, more specifically, an object detection problem, and that cannot be more true due to the simple fact that our primary input is images, this has put CNN as the major choice to deal with the problem, however recent studies use the fact that a text is a sequence of characters and start dealing with the problem with sequence to sequence dedicated architectures.

For the Arabic text, many studies have used LSTM, Bi-LSTM combined with CTC loss function to push the stat of the art (Graves et al., 2009) (Yousfi et al., 2015b) (Zayene et al., 2017). However, to our best knowledge, no studies has focused on using any encoder, decoder or even an encoder only structures for ASTR.

The main idea of this work was to find the best way to use an encoder decoder structure on the ASTR problem. We decide to go with the Transformer architecture because it combines the power of attention and self-attention

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

mechanism along with the encoder-decoder structure. We did may changes on many layers of the original transformer. starting from the simple embedding layer and ending with the best model optimizer.

Our best model pipe line is described in the pipe line section and illustrated in Figure 2.

Our method has proved to outperform the state-of-the-art accuracy, loss, and training / prediction time. All the results and the executions scenarios including those with bad results are discussed heavily in the experiment and results section.

4. PIPE LINE

4.1. Motivation

Recurrent neural networks, Long short term memory (Hochreiter & Schmidhuber, 1997) and any other sequence treating models has significantly affected the results obtained when dealing with any sequential problem. However big red flags appear on their ability to handle long range sequences. not to mention their mono execution nature which affect heavenly the training time of any sort of networks.

The Transformer is the natural successor of all these architectures because of its ability to deal with long sequence without suffering from memory loss on long sequences. It was introduced for the first time by (Vaswani et al., 2017) as an NLP dedicated architecture. What makes the transformer special is that it only uses the attention mechanism dispensing any kind of recursive behavior entirely.

From a global perspective, the transformer receives input data, pass it threw a linear mapping layer. positional information is than added to the results of the linear layer and forwarded to the encoder block which serves as memory holder and a second input for the decoder layer. The pipe line of standard Transformer is described in Figure 3.

4.2. Embeddings

Although an image in its format is processable by any ANN, feeding an image in its natural shape to the transformer will lead to a very large complexity when reaching the multi head attentions blocks. proposed the Vision transformer (ViT). They use only the encoder block of the transformer for image classification. In their work. they propose to divide the image into small patches then feed it to the network. More recent studies use the same architecture but ignore the image patching. they feed the Convolutional neural networks (CNN) results (often referred to as feature maps) to the encoder block and abandoned the decoder part entirely. With the goal of converting every input image to N vectors of dimension d_{model}, the C, H, W, image is first patched into N, C, 16*16 patches than mapped threw a learnable linear mapper to N, d_{model} vectors. In our study we use both image patching and CNN feature maps as an input of the encoder. each case has been studied as a stand-alone hypothesis and re results are discussed in the results chapter.

4.3. Positional encoding

The Transformer original paper propose the use Sinus and Co-sinus functions to add a positional information to the data before passing it to the next layers. The community has reported many positional encoding (PE) ways including learned and fixed PE. We experimented the behavior of our model when using different PE techniques and the learned PE matrix was the best in our case. We use $max_{patchesNumber} * d_{model}$ matrix of trainable randomly initialized parameters and add it to the embedding matrix.

4.4. The encoder

"The encoder is composed of a stack of N = 6 identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network" [@4531750]. We adapted the same encoder to our case. The challenge was to find the right number to use in order to increase the accuracy of the model while keeping its performance indicators in an accepted value

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.5. The decoder

The decoder block is composed of a stack of N = 4 identical layers. The first is a masked multi head self-attention layer. The mask prevents the networks from using information yet to come and let it use only the part that it is in the current scope of a given time t. This masking mechanism proved to be useful in many cases, but for our study we make the masking matrix learnable in order to simulate the human brain reading mechanism. In another words, we want the network to be able to use the beginning and the end of current input in order to recognize a char in the middle of the image. The rest of the decoder is employed as it is in the original structure

4.6. Parameters

The behavior of ANNs is highly impacted by the hyper parameters of the model. In our case, the model is based on many parameters. Using only 15 percent from the data, we did many heuristically guided changes to find out the best value for each one. The best value of the d_{model} was 256 which led to the same accuracy and significantly reduced the training time. On the other hand, using small values for the d_{ff} has affected the model accuracy negatively and we ended using 1024. Another interesting result was the fact that using more than 8 heads on the multi head attention block had no impact on the accuracy and just increased the training time. The reset of the parameters is described in table 3.

Parameter	Description	Best value
d_model	The model dimension	256
d_ff	The Feed forward mapper dimension	1024
N	Number of layers in the encoder stack	4
dropout	Dropout used across the model	0.1
h	Number of heads	8
Lr	The optimizer learning rate	104

Table 3 Model Parameters best values

5. EXPERIMENT AND RESULTS

5.1. Training

We used three different datasets, The first is ALiF data set, it contains 4152 images dedicated for training presupposes and 1132 for test. It is publicly available for no commercial use. The second dataset is the AcTiV-R dataset, it contains 10,415 cropped text images. The third is our dataset Iyad, it is described heavily it Section 5. The training has been done on 12 threads CPU and a two NVIDIA RTX 3060 graphic cards.

5.2. Metrics

For a better understanding of the system, many metrics has been used to evaluate and tune the system during the training and prediction steps. The loss calculation has been done using the The Kullback-Leibler divergence (KLDiv) loss following. where y_{pred} is the prediction matrix and y_{true} is the encoded representation of the correct label of the image

$$KLDiv = L(y_{pred}, y_{true}) = y_{true} * log \frac{y_{true}}{y_{pred}} \qquad \textit{Equation 1 Kullback-Leibler divergence}$$

To measure the accuracy of model, we use two commonly used metrics: Char recognition (CR) rate and Text recognition rate (TR) following recognized Equations respectively:

$$CR = \frac{correctlyRecognizedChars}{Totalchars} \qquad \textit{Equation 2 CR}$$

$$TR = \frac{correctlyRecognizedTexts}{TotalImages} \qquad \textit{Equation 3 TR}$$

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.3. Results

Analyzing the model behavior on each dataset show that the model is doing extremely well when enough data provided but overfitting when no data is provided. Training the model on the white backgrounds only from IYad dataset has very poor accuracy even when using the same type of images when calculating the accuracy. The only way the model gets to learn was when using all three types in the dataset. It is worth to mention that the model was not impacted by the size of the data set but by the diversity of input types. Full details about the data impact are illustrated on Table 4

	White + Noisy		**All**	
Size	50%	100%	50%	100%
Val loss	1.8	0.9	1.1	0.15
Trai loss	0.02	0.0089	0.0014	0.0032
Val Accuracy	87.9%	91.2%	86.3%	92.7%

Table 4 Data effect on model behavior from Iyad Dataset

Method	**ALiF**		**IYad**	
	Loss	**Accuracy**	**Loss**	**Accuracy**
CNN Transformer	2.285	61.3 %	0.96	86.7 %
Image patching Transformer	4.13	43.7 %	**0.04**	**91.2 %**
LSTM CTC	1.2	83.2 %	1.11	86.3 %

Table 5 Result from all execution scenarios

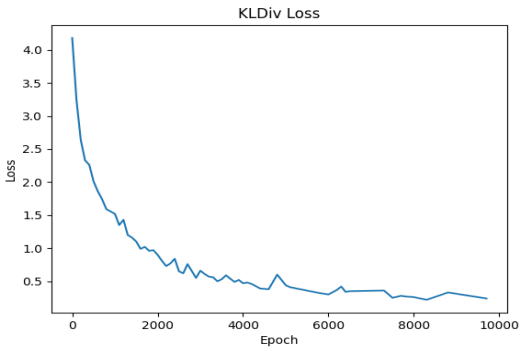

Figure 4 CNN + Transformer First loss

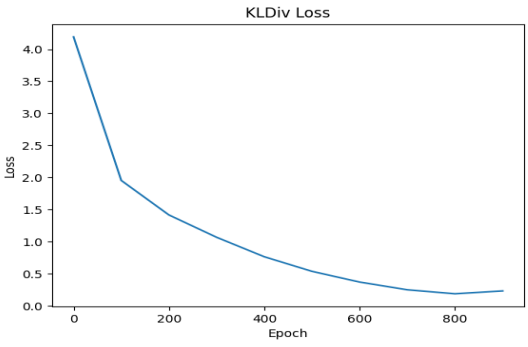

Figure 5 Image patching + Transformer First epoch loss

6. conclusion

We have presented in this paper a system for Arabic text recognition based on Transformer architecture. The system combines the original Transformer architecture and the patching technique from Vision transformer. We did some changes on the architecture itself and run many execution scenarios to find the best parameters for our subject. We

run the best model on two Datasets: ALiF and Iyad. We reported an accuracy on single fonts of 93.5 percent outperforming the state of the art.

This paper also introduced video data set we name it IYad. It is publicly available for scientific research purposes. The dataset is composed of two parts: Iyad-W, and Iyad-S. Each part contains 1,400,000 image and more are being generated. It is published with a set of tools to help maintaining it on the official web site of Mustapha STAMBOULI university. The data set is artificially created and manually content labeled. We aim to keep the DS updated and compatible with as many techniques as possible. A set of tools is published with the dataset to make manipulating and enhancing the dataset a strait forward process.

REFERENCES

[1] Al-Ohali, Y., Cheriet, M., & Suen, C. (2003). Databases for recognition of handwritten Arabic cheques. *Pattern Recognition*, *36*(1), 111–121. https://doi.org/10.1016/S0031-3203(02)00064-X

[2] Anthimopoulos, M., Gatos, B., & Pratikakis, I. (2007). Multiresolution text detection in video frames. *VISAPP (2)*, 161–166.

[3] Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., & Zhou, S. (2017, October). Focusing Attention: Towards Accurate Text Recognition in Natural Images. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[4] El Abed, H., Kherallah, M., Märgner, V., & Alimi, A. M. (2011). On-line Arabic handwriting recognition competition: ADAB database and participating systems. *International Journal on Document Analysis and Recognition (IJDAR)*, *14*, 15–23.

[5] Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(5), 855–868. https://doi.org/10.1109/TPAMI.2008.137

[6] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation*, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] Jaiem, F. K., Kanoun, S., Khemakhem, M., El Abed, H., & Kardoun, J. (2013). Database for Arabic Printed Text Recognition Research. In A. Petrosino (Ed.), *Image Analysis and Processing – ICIAP 2013* (pp. 251–259). Springer Berlin Heidelberg.

[8] Mezghani, A., Kanoun, S., Khemakhem, M., & Abed, H. E. (2012). A Database for Arabic Handwritten Text Image Recognition and Writer Identification. *2012 International Conference on Frontiers in Handwriting Recognition*, 399–402. https://doi.org/10.1109/ICFHR.2012.155

[9] Oualid, K., Fatma, B., & Rochdi, B. B. (n.d.). *IYaD: A dataset for Arabic printed text recognition in natural scene videos*. International Conference on Computer Science, Technology and Artificial Intelligence (ICCSTAI-25).

[10] Panhwar, M. A., Memon, K. A., Abro, A., Zhongliang, D., Khuhro, S. A., & Memon, S. (2019). Signboard Detection and Text Recognition Using Artificial Neural Networks. *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 16–19. https://doi.org/10.1109/ICEIEC.2019.8784625

[11] RAMDAN, J., Omar, K., Nasrudin, M. F., & Mady, A. (2013). Arabic Handwriting Data Base for Text Recognition. *Procedia Technology*, *11*. https://doi.org/10.1016/j.protcy.2013.12.231

[12] Shivakumara, P., Phan, T. Q., & Tan, C. L. (2011). A Laplacian Approach to Multi-Oriented Text Detection in Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 412–419. https://doi.org/10.1109/TPAMI.2010.166

[13]Slimane, F., Ingold, R., Kanoun, S., Alimi, A. M., & Hennebert, J. (2009). A New Arabic Printed Text Image Database and Evaluation Protocols. *2009 10th International Conference on Document Analysis and Recognition*, 946–950. https://doi.org/10.1109/ICDAR.2009.155

[14]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*.

[15]Wang, K., Babenko, B., & Belongie, S. (2011). End-to-end scene text recognition. *2011 International Conference on Computer Vision*, 1457–1464. https://doi.org/10.1109/ICCV.2011.6126402

[16]Yousfi, S., Berrani, S.-A., & Garcia, C. (2015a). ALIF: A dataset for Arabic embedded text recognition in TV broadcast. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 1221–1225. https://doi.org/10.1109/ICDAR.2015.7333958

[17]Yousfi, S., Berrani, S.-A., & Garcia, C. (2015b). Deep learning and recurrent connectionist-based approaches for Arabic text recognition in videos. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 1026–1030. https://doi.org/10.1109/ICDAR.2015.7333917

[18]Zayene, O., Amamou, S. E., & BenAmara, N. E. (2017). Arabic Video Text Recognition Based on Multi-dimensional Recurrent Neural Networks. *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 725–729. https://doi.org/10.1109/AICCSA.2017.126

[19]Zayene, O., Masmoudi Touj, S., Hennebert, J., Ingold, R., & Essoukri Ben Amara, N. (2018). Open Datasets and Tools for Arabic Text Detection and Recognition in News Video Frames. *Journal of Imaging*, *4*(2). https://doi.org/10.3390/jimaging4020032