Synthesizing versus Augmentation for Arabic Word Recognition with Convolutional Neural Networks

Reem Alaasam
Ben-Gurion University of the Negev
Beer-Sheva, Israel
Email: rym@post.bgu.ac.il

Berat Kurar Barakat
Ben-Gurion University of the Negev
Beer-Sheva, Israel
Email: berat@post.bgu.ac.il

Jihad El-Sana
Ben-Gurion University of the Negev
Beer-Sheva, Israel
Email: el-sana@cs.bgu.ac.il

***Abstract*—In this paper, we present a sub-word recognition method for historical Arabic manuscripts, using convolutional neural networks. We investigate the benefit of extending training set with synthetically created samples in comparison to augmentation. We show that annotating around ten pages of a manuscript and extending it, is sufficient for successful sub-word recognition in the whole manuscript. In addition, we show the contribution of using different combinations of training sets and compare their sub-word recognition performance in the whole manuscript.**

I. Keywords

Database, Arabic, handwritten, text recognition.

II. Introduction

Historical handwritten documents contain important information for scholars to study. In order to access the contents of historical documents, they are represented as digital images. While for printed documents Optical Character Recognition (OCR) systems reached high recognition rates, recognition in historical documents is still a subject of improvement. This task is challenging because of the degraded images adding to the cursive nature of Arabic scripts. Segmenting a word in Arabic handwritten historical document to letters to determine the query word is very difficult and does not provide good recognition rates. Instead, recognizing sub-word in a holistic approach provides better results, as it avoids the error-prone letter segmentation procedure. Most approaches for word recognition rely on various hand crafted features [1] of the text image. Such features achieve very good accuracy on clean images in contrast to distorted images.

Recently Convolutional Neural Networks (CNN) have shown great solutions for many visual task problems such as text recognition [2], image recognition [3] and character recognition [4]. CNNs input raw image and learn representation of the image through the convolutional layers and classify the image through the fully connected layers. Despite having better performance CNNs are harder to train in terms of training set size. The training set should be sufficiently large to represent the input space well. To prevent overfitting, the training set size should be at least several times [5] the total number of weights which may reach to tens of millions [3]. In case of historical documents it is difficult to have a dataset in sufficient size due to the required annotation effort. Providing annotations in most of the cases requires domain expertise, and is done manually. Hence to utilize a CNN for Arabic handwritten sub-word recognition, the training set size needs to be increased. Ahmad and Fink [6] investigated the use of computer generated text in different typefaces for setting up large amount of training data in handwritten Arabic text recognition.

This paper explores minimizing the manual annotation required to recognize the text of the whole manuscript with high accuracy rate, using two different dataset extension methods, synthesizing and augmentation. We created three training sets. The first training set is extracted from ten pages of a historical manuscript and called original training set. The second training set is the extension of original training set by synthesizing using the method in [7]. The third training set is the extension of original training set by augmentation. We analyzed the impact of using these training sets as well as increasing the size of each training set gradually, with a shallow CNN. All the three training sets are comparable in performance. Original and augmented data in combination outperformed the other two training sets. However synthesized data gives a good sense of original data to the network, but does not provide measurable improvement in performance over the original data.

In the rest of the paper, we briefly overview related work in Arabic text recognition in Section III; explain preprocessing, augmentation, and synthesizing of the dataset in Section IV; present network architecture, training detail and results of the experimental study in Section V; and finally in Section VI we summarize the paper and draw conclusions.

III. Related Work

During the past two decades Arabic text recognition has attracted the interest of researchers and many quality papers have been published. Early works deal with Arabic characters as isolated individuals. Among the early works an approach extracts outer contour or the skeleton of the characters to obtain Fourier descriptors [8], [9], [10] or utilizes the Fourier coefficients of the handwritten dynamic representation [11]. Another approach is Bayes classification using the class conditional density functions of Arabic characters [12]. Some segmentation based methods separate words into characters based on their geometrical and topological properties. [13], [14]. Other approaches segment the words into characters by vertical projection and histogram techniques [15], [16].

Segmentation was also made based on HMM models [14] or morphological rules which are constructed at the feature extraction phase [17].

The recursive script recognition is lead by segmentation free methods. Maddouri and Amiri [18] introduced global features specific to Arabic and rate the recognition system by propagating these features into a transparent neural network. Saabni [19] avoided segmenting words into individual letters by a multi-level recognizer for online Arabic handwriting in a holistic fashion.

Recent document processing algorithms extract interest points from gray scale images [20] and utilize these points for various applications, such as word spotting [21], [22] and writer identification [23]. Most of these algorithms control the distribution of feature points by imposing a grid or defining patches [24], [21]. The size of this grid and the number of sample points is defined in an ad-hoc manner. These algorithms have the drop on the binary-prerequisite-based algorithms for gray scale images. Other works are based on bag-of-visual-words model, such as [20], [25], [26], [27]. The performance of these algorithms deteriorates as the degradation level increases [28]. These points are used to compare the similarity among the components under the assumption that they faithfully represent the processed text components.

IV. Dataset

In this study we used the VML-HD dataset [29], which includes fully annotated historical manuscripts in Arabic, some part of a page is shown in Figure 1. We focus this study on a randomly selected set of sub-words and gradually increased the number of pages we have used for training. We have found that we need 10 pages to reach acceptable recognition rate (above 90%) on the rest of the manuscript.

We extract sub-words from 10 pages, and we shall refer to this original set of sub-words as the ORG set. We extended the ORG set using two procedures: augmentation and synthesizing. The augmentation procedure generates new samples from a given image by applying various linear transformations, while the synthesis procedure use the ORG and generate a data-structure that guides the synthesis of new samples [7]. We shall refer to the augmented and synthesized sets as AUG and SYN, respectively. Each train set is used to create three train sets:Train1, Train2, and Train3 of varying sizes, as shown in Table I.

The test set, shown in Table I, is extracted from the whole rest of the original historical document and used to test the models trained on the three train sets. Although the number of classes is relatively small, some classes are not visually very distinct, as shown in Figure 2.

A. Preprocessing

As the dataset contains a variety of image dimensions, each image is resized to a square aspect ratio and a resolution of 100×100 pixels. We also converted images to grayscale and normalized the pixel values into $[0, 1]$ range. Some resultant images after resizing and gray scaling are shown in Figure 3.

TABLE I
NUMBER OF SAMPLES IN THE DATASET

	Train1	Train2	Train3	Test	#Classes
ORG	344	670	1230	4124	39
ORG+SYN	685	1137	1696	4124	39
ORG+AUG	685	1137	1696	4124	39

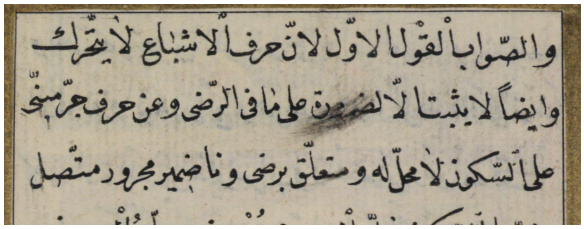

Fig. 1. A page part from the historical document that is used in the experiment.

B. Augmentation

The AUG set is created by augmenting the data samples in ORG with various transformations. Each sample is uniformly scaled in the range of $[0, 0.1]$, width shifted in the range of $[0, 0.1]$, height shifted in the range of $[0, 0.1]$ and rotated in the range of $[0, 20]$. An offline augmentation was used to generate multiple augmented image samples per training image. Some resultant images after transformations are shown in Figure 4.

C. Synthesizing

Automatic synthesis of historical handwritten Arabic sub-words is a novel framework that is proposed recently by [7]. It first builds a Letter Connectivity Map (LCM) (Figure 5) that includes multiple instances of each letter's various shapes, since an Arabic letter's shape varies by its position in the word. The LCM generated from around 10 annotated pages of the manuscript is then used to guide the automatic synthesis of Arabic sub-words that form the SYN set. The synthesized sub-words are picked by the user, for each sub-word if all the letters to build it exist in the LCM, then it can be synthesized.

Fig. 2. Example of visually similar sub-words.

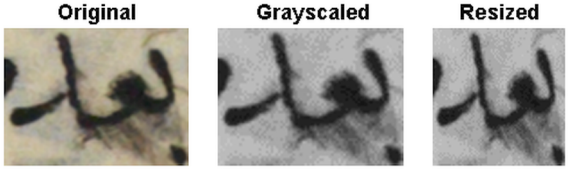

Fig. 3. Example of a preprocessed sub-word.

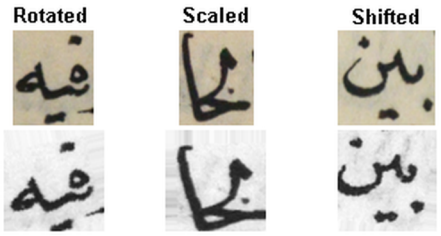

Fig. 4. Examples of augmentation.

Printed	ع
Isolate	ع ع ع
Initial	ع ع ع
Medial	ع ع ع
Final	ع ع ع

Fig. 5. An example for an LCM entry, and its output in various positions

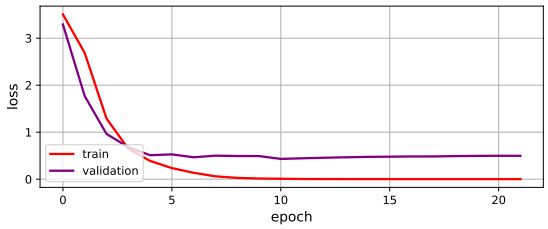

Fig. 6. 4-layer model loss on $ORG + AUG$.

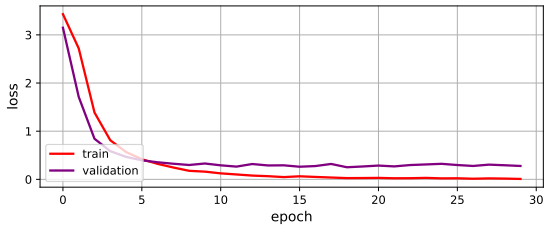

Fig. 7. 5-layer model loss on $ORG + AUG$.

V. Experimental Evaluation

In this experimental study we focus on the question how the synthesized and augmented data influence historical Arabic sub-word recognition using a CNN. We ran experiments on a 39 class dataset with the objective to compare the influence of different training sets on the test classification accuracy of the same CNN model. We started exploring with an architecture very similar to LeNet which is a well-studied architecture for classifying handwritten digits [30]. This was motivated by that our dataset consists of grayscale and non natural images.

A. Architecture

LeNet like model contains four layers, two convolutional layers and two fully connected layers. The convolutional layers have 20 and 30 number of filters with sizes 13×13 and 11×11 respectively. The fully connected layers have 500 and 39 neurons respectively. Convolutional layers include a ReLU nonlinearity, followed by 2×2 max-pooling. Stochastic Gradient Descent (SGD) with a learning rate of 0.01 is used to train the network.

During the experiments we noticed that the accuracy stopped at 1.0, and gave low bias in training sets, as shown in Figure 6. However the gap between the training and validation loss curves was due to high variance in validation set. In an effort to reduce the variance we added dropout with a rate of 0.5 and one more convolutional layer with 50 filters in size of 7×7 to conserve low bias. The resulting loss graph had smaller gap (Figure 7). This model achieved higher accuracy on the test set and it was used in all the experiments.

B. Training

The balance among the bias and the variance is improved by early stopping method. Each training set is further divided into two subsets. The first is called training set and used for computing the training loss and updating the weights. The second is called validation set and used to monitor the validation loss while training. Since the training sets have imbalanced classes, as depicted in Figure 8, we split them by a stratified 1-fold with a ratio of $75 - 25$. This provided two subsets with approximately the same class distribution as the training set.

During the initial phase of training the training error and the validation error decreases. However when the overfitting begins, the error on validation set begins to increase. Therefore after validation error increased for a patience of 10 epochs, we stopped training and picked the model with the best validation error rate.

C. Results

To explore the influence of data extension methods, we started by small train sets and then approximately doubled their sizes two times. While the sizes of train sets are increased from left to right as shown in Table II, we fixed the network as the LeNet architecture but with 5-layers .

Horizontally comparing the results in Table II, the best accuracies are achieved by Train3. Not surprising, since larger train set size provides better test accuracy.

If we compare the results in Table II vertically, the best test accuracies are achieved by ORG+AUG (the combination of ORG and AUG data sets). ORG and ORG+SYN achieved roughly the same. This means the network is able to learn synthesized images in the same way as the original images. However; there are some distortions, since ORG got better results than ORG+SYN, but since it was so close it does not seems that the distortions is very big. On the other hand; ORG+AUG achieved accuracy higher than ORG, which show

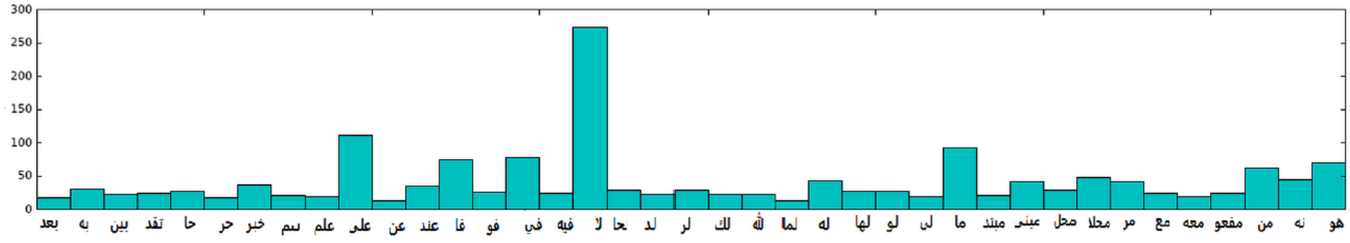

Fig. 8. Histogram of class sizes in original training set.

Fig. 9. Histogram of precision and recall on each class.

that increasing the dataset by augmentation seems to be better than increasing the size of the original dataset (Table I).

TABLE II
ACCURACY RESULTS OF ALL EXPERIMENTS

	Train1	Train2	Train3
ORG	0.9520	0.9677	0.9782
ORG&SYN	0.9340	0.9658	0.9748
ORG&AUG	0.9661	0.9840	0.9847

In addition to accuracy; We also present precision and recall per class. These matrices are important, since it is often preferred to reduce the number of false negatives, which is when a sub-word is falsely classified to another class, compared to the number of false positives, which is when another sub-word is falsely classified to the intended class. In practical applications false negatives will go unrecognized, but false positives will only need an additional look to confirm the target sub-word.

The best performing model's precision and recall values for each class is depicted in Figure 9. The sub-words في and علم achieved the lowest precision rates. This was due to the small number of test samples which were 1 and 3, respectively. But the number of false negatives in these two classes was 0 and they achieved a high recall rate as usually preferred in word recognition applications. The sub-word عن achieved the lowest recall rate. We saw in the confusion matrix that this sub-word was falsely predicted as من two times and as مر one time, presumably because these two sub-words are visually similar to the sub-word عن (Figure 10).

Fig. 10. Low recall rate for the sub-word عن was presumably due to its visual similarity with other two sub-words.

VI. CONCLUSION

In this paper we explored the benefit of using synthesized and augmented data for historical Arabic sub-word recognition. Using ten pages of a historical manuscript we achieved impressive performance on word recognition in the whole rest of the historical manuscript. We ran experiments on varying amounts of original, synthesized and augmented train sets. Our main findings are: 1) Using ten pages of a manuscript is sufficient for successful word recognition in the rest of the manuscript. 2) Extending the data set by augmentation is better than adding synthesized data. 3) Synthesized images are similar to the original images, but with a bit of distortions, however; the network is able to learn the information in them.

ACKNOWLEDGMENT

The authors would like to thank the Lynne and William Frankel Center for Computer Science.

REFERENCES

[1] A. Khémiri, A. K. Echi, A. Belaïd, and M. Elloumi, "A system for off-line arabic handwritten word recognition based on bayesian approach," in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE, 2016, pp. 560–565.

[2] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3304–3308.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng, "Text detection and character recognition in scene images with unsupervised feature learning," in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 440–445.

[5] B. Widrow and M. A. Lehr, "30 years of adaptive neural networks: perceptron, madaline, and backpropagation," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1415–1442, 1990.

[6] I. Ahmad and G. A. Fink, "Training an arabic handwriting recognizer without a handwritten training data set," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 476–480.

[7] M. Kassis and J. El-Sana, "Automatic synthesis of historical arabic text for word-spotting," in *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*. IEEE, 2016, pp. 239–244.

[8] S. A. ALshebeili, A. A.-F. Nabawi, and S. A. Mahmoud, "Arabic character recognition using 1-d slices of the character spectrum," *Signal Processing*, vol. 56, no. 1, pp. 59–75, 1997.

[9] T. S. El-Sheikh and R. M. Guindi, "Automatic recognition of isolated arabic characters," *Signal processing*, vol. 14, no. 2, pp. 177–184, 1988.

[10] S. A. Mahmoud, "Arabic character recognition using fourier descriptors and character contour encoding," *Pattern Recognition*, vol. 27, no. 6, pp. 815–824, 1994.

[11] N. Mezghani, A. Mitiche, and M. Cheriet, "On-line recognition of handwritten arabic characters using a kohonen neural network," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*. IEEE, 2002, pp. 490–495.

[12] ——, "Bayes classification of online arabic characters by gibbs modeling of class conditional densities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1121–1131, 2008.

[13] H. Almuallim and S. Yamaguchi, "A method of recognition of arabic cursive handwriting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 5, pp. 715–722, 1987.

[14] A. Amin, A. Kaced, J. Haton, and R. Mohr, "Handwritten arabic character recognition by the irac system," in *Proc. 5th Int. Conf. Pattern Recognition*, 1980, pp. 729–731.

[15] A. Amin and J. F. Mari, "Machine recognition and correction of printed arabic text," *IEEE Transactions on systems, man, and cybernetics*, vol. 19, no. 5, pp. 1300–1306, 1989.

[16] K. El Gowely, O. El Dessouki, and A. Nazif, "Multi-phase recognition of multifont photoscript arabic text," in *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, vol. 1. IEEE, 1990, pp. 700–702.

[17] T. Sari, L. Souici, and M. Sellami, "Off-line handwritten arabic character segmentation algorithm: Acsa," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*. IEEE, 2002, pp. 452–457.

[18] S. S. Maddouri and H. Amiri, "Combination of local and global vision modelling for arabic handwritten words recognition," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*. IEEE, 2002, pp. 128–135.

[19] R. Saabni and J. El-Sana, "Hierarchical on-line arabic handwriting recognition," in *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ser. ICDAR '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 867–871.

[20] M. Rusinol, D. Aldavert, R. Toledo, and J. Lladós, "Browsing heterogeneous document collections by a segmentation-free word spotting method," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 63–67.

[21] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Efficient Exemplar Word Spotting," in *British Machine Vision Conference*, 2012, pp. 67.1–67.11.

[22] A. Kovalchuk, L. Wolf, and N. Dershowitz, "A simple and fast word spotting method," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014, pp. 3–8.

[23] S. Fiel and R. Sablatnig, "Writer retrieval and writer identification using local features," in *In Proc. 10th International Workshop on Document Analysis Systems*, March 2012, pp. 145–149.

[24] V. Dovgalecs, A. Burnett, P. Tranouez, S. Nicolas, and L. Heutte, "Spot It! Finding Words and Patterns in Historical Documents," in *12th International Conference on Document Analysis and Recognition*, 2013, pp. 1039–1043.

[25] R. Shekhar and C. Jawahar, "Word image retrieval using bag of visual words," in *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*. IEEE, 2012, pp. 297–301.

[26] J. Lladós, M. Rusinol, A. Fornés, D. Fernández, and A. Dutta, "On the influence of word representations for handwritten word spotting in historical documents," *International journal of pattern recognition and artificial intelligence*, vol. 26, no. 05, p. 1263002, 2012.

[27] K. Zagoris, I. Pratikakis, and B. Gatos, "Segmentation-based historical handwritten word spotting using document-specific local features," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014, pp. 9–14.

[28] I. Rabaev, I. Dinstein, J. El-Sana, and K. Kedem, "Segmentation-free keyword retrieval in historical document images," in *International Conference Image Analysis and Recognition*. Springer, 2014, pp. 369–378.

[29] V. M. L. at Ben-Gurion University of the Negev, "Arabic historical documents dataset," December 2016. [Online]. Available: http:www.cs.bgu.ac.il/vml

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.