2017 14th IAPR International Conference on Document Analysis and Recognition

# Alignment of Historical Handwritten Manuscripts using Siamese Neural Network

Majeed Kassis*
Computer Science Department
Ben-Gurion University of the Negev
Beer-Sheva, Israel
majeek@cs.bgu.ac.il

Jumana Nassour*
Computer Science Department
Ben-Gurion University of the Negev
Beer-Sheva, Israel
jumanan@cs.bgu.ac.il

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

***Abstract*—Historical manuscript alignment is a widely known problem in historical document analysis, and the attempt of finding the differences between manuscript editions is mainly done by hand. Today, most of the computational tools coming to assist the historians are based on word recognition or spotting. These solutions are partial at best. In this paper, we present a Siamese neural network based system, which automatically identifies whether a pair of images contain the same text without the need of recognizing the text. The user is required to annotate several pages of two manuscripts, and with the assistance of synthetically generated data and affine distortions we can align two manuscripts written by different writers, achieving strong results.**

## I. Introduction

Manuscript alignment is one of the important tasks in historical manuscript research. It aims to determine the similarities and differences between two versions of a given manuscripts, mostly written by different writers. Currently, this tiresome and time-consuming procedure is done manually. This dissimilarity between various copies of the same manuscript arose from the publishing/copying procedures of the past. Each time a manuscript is copied, scribes often omit, insert, or replace words to adapt the content to different geographical regions or eras. Sometimes, scholars perceive the copied version as their personal copy and they embed their own explanation and notes on the original manuscript into the copied version itself. Contemporary researchers study the differences between different versions of a manuscript and attempt to explain the reasons behind these differences to reveal the original content of the manuscript.

In this work we present a Siamese network based system for aligning historical handwritten manuscripts. The manuscripts were written by different writers resulting in different writing styles yet almost identical text, as exemplified in Figure 1. The Siamese Convolutional Neural Network assists the system in deciding whether a pair contains same text, or different text.

Siamese neural networks consist of two identical networks that accept two inputs and are joined by an energy function that computes a distance metric between the last hidden layer of the two identical networks. This structure ensures the consistency of the predictions made by the network, which is invariant to the order of the two fed images due to its symmetry. Siamese neural networks were first introduced to solve signature verification problems [1], and since then they have been used for other verification tasks, such as face verification.

Convolutional neural networks, which consist of several layers of convolutions, are excelling at learning generic features found in images. The purpose of a convolution operation is to provide direct filtering interpretation, where each feature map is convolved against input features to identify patterns in the input's pixel groupings. As a result, the output of each layer consists of the important spatial features. Since its first appearance in attempting to recognize handwritten digits [2], it gained extensive popularity, and it has been widely used in various image related domains providing state of the art results.

The remainder of the paper is organized as follows. In Section II we briefly overview of closely related work. In Section III we present the system overview. Next, in Section IV, we explain in detail the data preparation phase, a critical step in neural networks training, since it assists neural networks to converge and provide minimal loss with maximal accuracy. Then, in Section V, we explain the structure of the Siamese neural network we've used in our alignment algorithm, and explain in detail the alignment algorithm which uses the Siamese network to align the manuscripts. Finally, in Section VI, we discuss the experiments we've initiated for both training the network as well as the alignment algorithm. We conclude our work, which provides very strong results, with some insights to planned future work.

## II. Related Work

Alignment can be done on several levels, starting from character alignment, to sub-word alignment, to complete words and sentences. Manuscript text alignment may be classified into two main groups: (1) supervised alignment, (2) unsupervised alignment. In the first category, a model needs to be learned from annotated manuscripts, while the second category uses image features and transcription information to apply the alignment correctly. In supervised alignment, one of the recent works uses hidden markov models to align Latin [3] script, another uses a character recognition model and geometric

* authors contributed equally

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.56

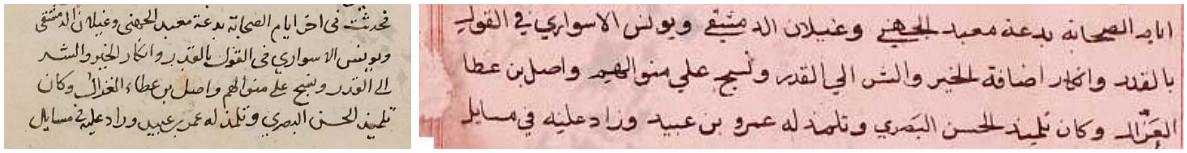


Fig. 1: Snippet of the manuscripts with different writing styles.

context for Chinese transcript mapping [4]. In unsupervised alignment, recent works align images of words with their text transcriptions in the context of building historical document retrieval systems [5]. Another work proposes an efficient transcript mapping technique to ease the construction of document image based on Hough transform guided text lines count [6]. Another method is guided by local and global features using a scoring system [7].

Convolutional neural networks has been first presented back in 1989s [2], but only recently has it begun to receive an ever increasing attention. Every year, new and better networks are being released [8], [9], [10], due to the release of datasets of massive sizes and the visual recognition challenges based on these datasets [11].

The interest in convolutional neural networks for handwritten documents has been growing in recent years, and many approaches are introduced every year. One work used a pre-trained convolutional neural network and then fine-tuned it to learn classes of word images [12]. Another work combined convolutional neural network with Conditional Random Field graphical model, taking the whole word image as a single input [13]. Another group presented a convolutional neural network architecture trained with embedded attributes to perform word spotting [14].

Siamese Neural Networks were first introduced to solve signature verification as an image matching problem [1]. A Siamese neural network consists of a network duplicated and joined at its end by an energy function. The function computes a metric to decide whether a pair of inputs is the same or not. The parameters between the cloned networks are tied and thus each image of the pair affects its weights, and since they are symmetric, it is invariant to the order of a pair. Multiple network architectures were presented and used for different tasks, such as face verification [15], image recognition [16], and object tracking [17].

## III. System Overview

To align two handwritten manuscripts we need a model to determine whether a pair of images, one taken from the first manuscript, and the other taken from the second manuscript, contain the same text. For this purpose, we train a Siamese convolutional neural network. This network architecture contains two or more identical subnetworks that share the same structure, parameters and weights, which are propagated across these subnetworks. As a result, these networks excel at finding similarities between given inputs.

We begin with data preparation phase, which is an important aspect in neural network training. Using this acquired data we train the neural network. Then, we detail the results achieved in the training phase, and test their validity using both validation and test sets that include data samples unseen by the network. We proceed by explaining the alignment system in which we are assisted by the model, which was generated from training the neural network, and the Hungarian algorithm to achieve strong results.

## IV. Data Preparation

In this section we detail the data preparation process. The creation of training, testing, and validation sets, and their structure. The main purpose of this process is to create a representative datasets to allow the neural network to converge properly, and without over-fitting.

### A. Ground Truth Generation

To train a Siamese neural network we need annotated data. We utilized WebGT [18] to annotate to the sub-word level several pages from two versions of the same manuscript. Roughly, 15 pages from each manuscript, where each page contains on average 270 sub-words. The obtained annotated sub-words form the basis of the training set. In addition, we annotated additional 36 pages from each manuscript, for validation and testing purposes. An illustration of one annotation example can be seen in Figure 2.

Once the data is annotated, we segment the sub-words using their bounding boxes (which are part of the annotated data). Since the bounding box used is rectangular, many of the images contain parts of other sub-words in addition to the sub-word originally annotated, as can be seen in Figure 3.

### B. Datasets Preparation

To enrich our training set and to assist the neural network to converge, we've used the recently released automatic synthesis system by [20]. The system uses the annotated pages, which are used for training, to generate more sub-word images. In addition, we further increase the number of samples by applying minor affine distortions. For each sub-word image, we apply an affine transformation, stochastically determined by a multidimensional uniform distribution, in the same manner mentioned in [16].

Due to the use of the fully connected layers in our network, we are unable to provide the neural network inputs of various sizes. Thus, we are required to scale the images to fit a single size. To reduce the distortion made by naive resizing, we scan the dataset and find the image with the largest height. Using this height, we scale, keeping aspect ratio, the images to the new height. Then, we calculate the median width of the set,

Fig. 2: Snippet of the annotation manuscripts with almost identical text, written by two writers having different writing styles.

and resize the images to fit the new width, keeping the height intact.

### C. Pair Generation

We denote a pair of images, one taken from the first manuscript and another taken from the second manuscript, as *true-pair* if they contain the same text, and *false-pair* if they contain different text. The dataset generation aimed to ensure maximal randomization of the dataset and representativeness of the manuscript to avoid over-fitting, i.e., each sub-word has the same representation ratio in the true-pairs as well as in the false-pairs. True-pairing examples can be seen in figure 3.

*1) Training Set:* To generate true-pairs for the training set, we look up for each sub-word its instances in the first manuscripts and its corresponding instances in the second manuscript and generate all possible pairs. To make the dataset representative as possible, we took the same number of true-pairs for each sub-word form. For the first training set we randomly chose 100 true-pair samples for each sub-word form, totaling $57,758$ pairs. For the second training set we randomly chose 500 true-pair samples for each sub-word form, totaling $101,537$, and for the last dataset we chose $1,000$ pairs for each sub-word form, totaling $151,226$ pairs.

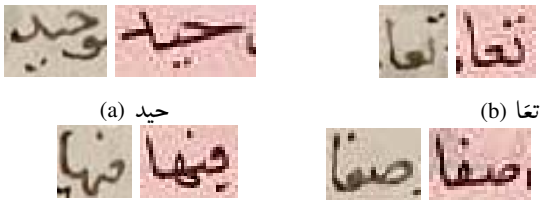
To generate the false-pairs for the training set, we pair for each image of a sub-word present in the true-pairs sub-set a randomly chosen image containing different text taken from the other manuscript. This method ensures providing the neural network for each image a true-pair and a false-pair example. This is done for all sub-words present in the true-pairs sub-set for both manuscripts. As a result the training sets are doubled in size to $115,516$, $203,074$, and $302,452$ respectively.

*2) Validation and Test Sets:* To generate both validation and test sets we use the sub-words taken from the manually annotated 36 pages of each manuscript, totaling 72 pages. Both the validation and test sets do not contain synthetic or affine distorted images, only manually annotated sub-words, since the purpose of our neural network training is to test it on real handwritten documents.

The generation of the true-pairs and the false-pairs is done in the same manner we use to generate the pairs for the training set. Once generated, we split the set into two equally sized parts, half of them for the validation set, and the other half for the test set.

It is important to note that since the ground-truth data used for the training set and for the validation/testing sets is not the same, there is no single pair to be found in both training and validation/testing sub-sets. This means that the neural network does not see a pair taken from the testing or validation during training. This is done to ensure that erroneously trained networks, which reduced the loss due to over-fitting and not generalization, fail once tested on either the validation or the test sets.

These efforts resulted in $292,128$ pairs total. We've set $146,578$ pairs for the validation set, and $145,550$ pairs for the testing set.



(a) حيد (b) تعَا

(c) فيهَا (d) صفَا

Fig. 3: Sample of pairs generated as ground truth input for the Siamese neural network

## V. Manuscript Alignment

We opted to adopt a Siamese neural network which, once trained, can determine with high probability whether two sub-words, each taken from a different book written in different writing style contain the same text or not, without the need to identify their text. Using the trained model we implement a rule based algorithm using Hungarian method to decide on the alignment of the script.

### A. Siamese Neural Network

We base our system on the Siamese convolutional neural network presented in [16] with several modifications that aim to avoid over-fitting. Upon the end of the first epoch the training loss continued to decrease but the validation loss began to increase, and after four epochs only, the neural network training accuracy dropped severely while the training loss increased dramatically, in contrary to their results, where they trained the network for 200 epochs successfully. This behavior deemed their network, as is, unfit for our needs.

Our network follows the same rough structure with several modifications. We've added a dropout layer, and another fully connected layer in addition to the original fully connected layer. We also modified the convolutional layers to have different kernel sizes ranging from $5 \times 5$ down to $2 \times 2$ and modified the number of features per layer to begin with 64, and multiplying their count by two, up to 512 features for the last convolutional layer. These modifications aim to overcome two problems: (1) the over-fitting problem due to input difference, and (2) the input image dimension difference, which motivated

us to modify the kernel sizes to achieve a meaningful result. An illustration of the network architecture can be seen in Figure 4. Of course, just like the neural network we based our work on, we calculate the absolute linear distance between the two twins, and finalize it by using a sigmoid function.

### B. Manuscript Alignment

The alignment algorithm uses a window of a fixed size, which moves along the document line one sub-word at a time. Each window contains several sub-words of both manuscripts. Let $w_1$ be $sb_1^1 ... sb_n^1$ and $w_2$ be $sb_1^2 ... sb_n^2$, we process the two windows using the trained model of the Siamese convolutional neural network detailed above, and calculate the absolute linear distance between each two sub-words that the neural network decided that they are a true-pair. Then, using the Hungarian method we find the most optimal fit possible for the given window while maintaining a minimal number of fit intersections between the two windows and following one restriction: if $sb_i^1$ is paired with $sb_j^2$ as best fit, it is not possible for $sb_{i+l}^1$ where $l > 0$ to be paired with any $sb_k^2$ where $k < j$. If this case happens, we apply Hungarian method again while excluding this pairing possibility from being an option. Once we get a proper pairing, we move onto the next window, and apply Hungarian method again after computing the absolute linear distances for every two sub-words in the new window using our trained model.

Since the windows overlap, each time the result obtained in a specific sliding window is the same as the previous one for a specific sub-word, the pairing confidence increases. In cases where the pairing stays the same, the confidence of the pairing increases. In cases where it contradicts the confidence is reduced. Once the confidence is reduced below a certain threshold, it becomes uncertain, and we decide that such a pairing is incorrect and exclude it. This iterative process is done until we reach end of the line. Since the manuscripts are already segmented, the sub-words can be arranged in one long line for the complete process.

## VI. Experiments

In this section we will explain the types of experiments we've conducted on both, the Siamese neural network as well as the alignment system. We will begin with detailing the experiments we've conducted on the neural network which achieved high accuracy on several input sets, and then we will move to the alignment system which used the trained neural network to align two different manuscripts successfully.

### A. Siamese Neural Network

After the data generation phase, and the creation of the training, testing, and validation sets. We initialized our network weights in the same manner mentioned in [16], both the convolutional layers and the fully connected layers are initialized from a normal distribution with zero-mean. We trained the network using the training sets of different sizes on a single nVidia 1080GTX. We've used Keras [21] front-end and Theano [22] back-end. The input is a pair of gray-scale images of size $83 \times 69$. We ran the training over $200$ epochs, after each one, the neural network performance was tested on the validation set, of size $146,578$, and once the training was complete we tested it on the test set, of size $145,550$.

We tested the neural network on different training set sizes, while keeping the validation and test sets the same. We chose three sizes of training sets, each time containing up to a specific number of each sub-word pairs randomly chosen from all possible sub-word pairs: up to $100$ pairs, up to $500$ pairs, and up to $1,000$ pairs. For the first option, the training set contained $115,516$ pairs, $203,074$ pairs, $302,452$ pairs, respectively. We trained the neural network over $200$ epochs for every training sub-set input size, and chose the model of the epoch providing the best result for each set. For complete results of these experiments, please refer to Table I.

| Training Set Size | Average Epoch Training Time | Validation Set Accuracy | Test Set Accuracy | Best Epoch |
|---|---|---|---|---|
| 115,516 | 695s | 96.36% | 96.39% | 42 |
| 203,074 | 1099s | 97.10% | 97.09% | 29 |
| 302,452 | 1533s | 97.41% | 97.36% | 22 |

TABLE I: Training and accuracy results following three data inputs.

As it can be seen in Table I, the more data we feed the neural network the better its accuracy results over the validation and test sets are. Since both the validation data, as well as the test data are not present in the training sets, we can conclude that the presented Siamese neural network has succeeded in achieving generalization. For the complete network architecture please refer to Figure 4.

In addition to the ground truth data mentioned above, we annotated two additional manuscripts taken from a different dataset and written by two new writers, for the purpose of testing the network generalization success. From the first manuscript we annotated $11$ pages containing on average $280$ sub-words, and from the second manuscript we annotated $20$ pages containing on average $170$ sub-words. We've generated the training, validation, and test sets in the exact same manner as mentioned in the previous section, resulting in training set sizes of $168,938$, and $397,332$. For the validation and testing sets, we've annotated additional pages of roughly the same quantity, resulting in $105,284$ pairs for the validation set, and $104,988$ pairs for the test set. These validation and testing sets, as before, do not contain synthetically generated pairs nor affine distorted pairs. The two additional manuscripts were not used in the alignment process, but only to test the network's ability to detect whether a pair of images is a true-pair, or a false-pair.

We've trained the neural network on these training sets, using the same settings as before, over $200$ epochs and chose the model of the epoch providing the best result for each input set. For complete results of these experiments, please refer to Table II.

We successfully trained the neural network separately on the first two manuscripts, and on the second two manuscripts. These four manuscripts are all written in different writing

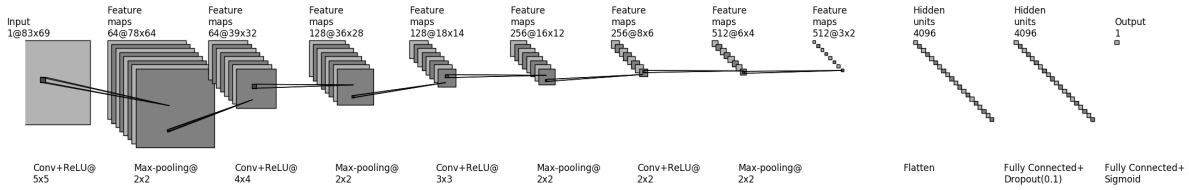



Fig. 4: Siamese convolutional neural network architecture. The Siamese twin connects after the last fully connected layer, where L1 distance is calculated from the feature vectors and the result is provided by the Sigmoid function.

| Training Set Size | Average Epoch Training Time | Validation Set Accuracy | Test Set Accuracy | Best Epoch |
|---|---|---|---|---|
| 168,938 | 799s | 96.33% | 96.31% | 35 |
| 397,332 | 1912s | 96.75% | 96.73% | 80 |

TABLE II: Training and accuracy results following three data inputs of the two additional manuscripts.

styles, thus, our network architecture has succeeded in generalization over both sets of manuscripts, achieving remarkable results.

### B. Manuscript Alignment

We tested the alignment algorithm over the first two manuscripts, we've used the 36 annotated pages of each manuscript (72 pages total). These same pages were used in the testing and validation of the Siamese neural network and have provided an accuracy of $97.41\%$ over the validation set. We've applied the alignment algorithm using a sliding window of size 9, each time applying the Hungarian method and moving the window one sub-word ahead. Following the algorithm, we've reached sub-word alignment accuracy of $94.73\%$.

It is true that the model has received higher accuracy rating on the same data than the complete system. This is due to words present in the first manuscript being replaced with other words in the second manuscript. These words and their replacement can share some sub-words, so the model will return true for these pairs (which is technically correct), but in our manual alignment they are not considered a pair. For such an example, please refer to Figure 5.

Please refer to [23] in order to download the data of the first two manuscripts, and refer to [24], under dataset section, to access the data for the second pair of manuscripts.

## VII. Conclusion and Future Work

In this work we've presented a novel system for historical handwritten manuscript alignment. The system takes as input several annotated pages from the two books we wish to align. Using these annotated pages we generate synthetically more data to train the neural network. We train a model to decide whether a pair of images, each taken from a different manuscript, contain identical or different text. Assisted by the model, we can align the rest of the two manuscripts identifying differences between them, as a result. From the experiments we've seen that the neural network has succeeded and the system as a whole has achieved impressive results.

A possible improvement we wish to explore, is to train the system on the word level, instead of the sub-word level. Since in cases where words are replaced, sometimes they share several sub-words which might cause the algorithm to provide incorrect result, and forces us to either detect the word from given sub-words.

Also, until now, the training of the system encapsulates the prior knowledge that the input is specific to two writers only, which is done by the training input itself. In the future, we will study the possibility of training the system to identify whether a pair contains the same text regardless of the writing style by training it on several writing styles and checking the possibility of generalization.

Finally, we aim to optimize the system to reduce the manual annotation requirement even further, which can boost the usability of the system even more.

## VIII. Acknowledgment

We would like to thank *Hussien Othman* for annotating the manuscripts and creating the ground truth information.

## References

[1] Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *IJPRAI*, 7(4):669–688, 1993.

[2] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 396–404. Morgan-Kaufmann, 1990.

[3] Andreas Fischer, Volkmar Frinken, Alicia Fornés, and Horst Bunke. Transcription alignment of latin manuscripts using hidden markov models. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pages 29–36. ACM, 2011.

[4] Fei Yin, Qiu-Feng Wang, and Cheng-Lin Liu. Transcript mapping for handwritten chinese documents by integrating character recognition model and geometric context. *Pattern Recognition*, 46(10):2807–2818, 2013.

[5] Svitlana Zinger, John Nerbonne, and Lambert Schomaker. Text-image alignment for historical handwritten documents. In *IS&T/SPIE Electronic Imaging*, pages 724703–724703. International Society for Optics and Photonics, 2009.

[6] Nikolaos Stamatopoulos, Georgios Louloudis, and Basilis Gatos. Efficient transcript mapping to ease the creation of document image segmentation ground truth with text-image alignment. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 226–231. IEEE, 2010.

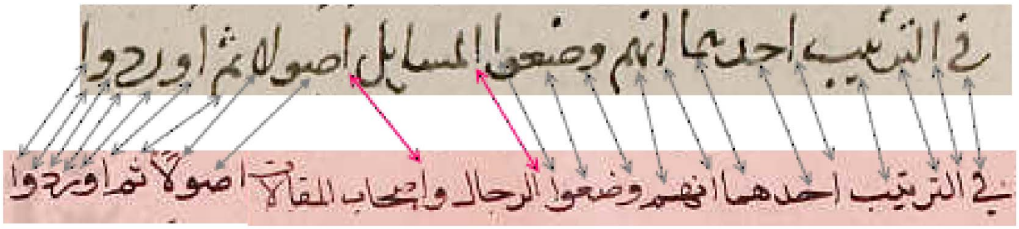


Fig. 5: Gray arrows denote correct alignment, while red arrows denote wrong alignment due to words being replaced with other words while sharing some of the sub-words.


[7] Nikolaos Stamatopoulos, Basilis Gatos, and Georgios Louloudis. A novel transcript mapping technique for handwritten document images. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 41–46. IEEE, 2014.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[10] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[12] Arjun Sharma et al. Adapting off-the-shelf cnns for word spotting & recognition. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 986–990. IEEE, 2015.

[13] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep structured output learning for unconstrained text recognition. 2014.

[14] Sebastian Sudholt and Gernot A Fink. Phocnet: A deep convolutional neural network for word spotting in handwritten documents. *arXiv preprint arXiv:1604.00187*, 2016.

[15] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.

[16] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *Proceedings of the 32nd international conference on machine learning, ICML*, 2015.

[17] Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*, pages 850–865. Springer, 2016.

[18] Ofer Biller, Abedelkadir Asi, Klara Kedem, Jihad El-Sana, and Itshak Dinstein. Webgt: An interactive web-based system for historical document ground truth generation. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 305–308. IEEE, 2013.

[19] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[20] Majeed Kassis and Jihad El-Sana. Automatic synthesis of historical arabic text for word-spotting. In *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*, pages 239–244. IEEE, 2016.

[21] François Chollet. Keras, 2015.

[22] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. May 2016.

[23] Majeed Kassis, Alaa Abdalhaleem, Ahmad Droby, Reem Alaasam, and Jihad El-Sana. Vml-hd: The historical arabic documents dataset for recognition systems. In *1st International Workshop on Arabic Script Analysis and Recognition*. IEEE, 2017.

[24] Majeed Kassis. The VML Arabic Historical Documents Dataset for Recognition Systems. http://www.cs.bgu.ac.il/~majeek, 2016. [Online; accessed 2016].