Degraded Document Image Repair with Fusion of Deep Networks and Template matching

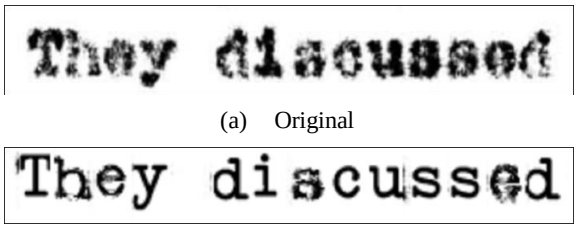
Pengtao Hu
University at Buffalo, SUNY
Buffalo, NY 14260, U.S.A.
pengtaoh@buffalo.edu

Abstract

Although the document scan techniques can digitize documents almost without loss of information for later recognition by OCR systems or directly presentation, the document images are required to be repaired in the cases that the document is so severely degraded that they are not recognizable by OCR system or readable by a human. We propose an approach to repair severely degraded document images using a fused method of deep neural networks and template matching. The basic idea is to leverage the deep neural network, which is with strong ability to classify variegated images, to make it possible to repair more severely degraded document images and to use template matching as verification and localization to obtain security as well as interpretation of repair process. We use two sequential deep convolutional neural networks (DCNN), one for segmentation and one for recognition, to generate reference location and patterns respectively of each patch that contains one character. Then according to the reference location and patterns, Euclidean distance transform based template matching generates hypotheses healthy versions of a patch. In the end, a voting mechanism takes these matched templates as voters is performed to determine repair result. Experimental results demonstrate the effectiveness of the proposed method.

1. Introduction

Degradations in document images can be caused by various factors including poor quality of paper, ink diffusion, and fading, errors happen in writing or printing process, inappropriate scanning, etc. The goal of document repair is to eliminate some of these degradations and generate an image that is closer to the ideal version that would be made under perfect conditions. Since documents are commonly used as a medium of information, the techniques of document repair can be applied to a variety of fields such as document recognition, historical document analysis, etc. The rapid digitization process of books [1] and other documents give rise to an imperative need for such repair techniques that could improve the readability of documents by human as well as recognition systems.



(a) Original

(b) Repaired

Figure 1: A part of a document image of target data and the result from the proposed method.

In the past decades, many efforts have been made and a large number of diverse methods have been proposed to address the problem of document image repair. Due to the diversity of degradation, the methods are usually proposed for certain kind of degradation e.g. [6] is specifically for bleeding-through or several kinds of degradation like binarization methods [3,4]. Besides, it is hard to evaluate a method precisely and completely since there is no sophisticated dataset for document image repair. Normalization methods are proposed to increase the contrast between foreground and background, e.g. [2]. Among all the methods, binarization or thresholding methods such as [3,4] account for a large proportion. These methods emphasize foreground and wipe off background by classifying them into two classes and usually assign them to be totally black and totally white respectively. Besides, there exist datasets commonly used by researchers and competitions on document image binarization [5]. Some other methods view this problem from a signal processing perspective. For instance, [6,7] use wavelet transfer to filter out background and interfering pattern from real strokes. Filters, detectors, and morphological operator are also applied in these methods e.g. [6,8].

As degradation of documents become much more variegated when it is severe, a robust method is required to repair the documents. Some machine learning methods bring out strong abilities to handle changeful inputs on this

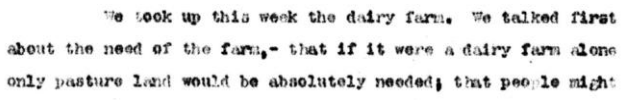
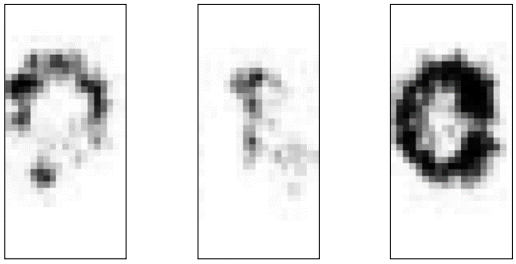


Figure 2: Several lines of an image in target data



(a) A '9' (b) A 't' (c) A 'c'

Figure 3: Some degraded characters in target data

problem. For instance, [9] learn the patterns in a document from the document itself by clustering and then model the document image as a Markov random field (MRF) on patches to enables the model to use context for repair. The method repairs various degradation including cuts, merges, blobs and erosions without supervised training. One of the requirements the data should satisfy for this method is that the healthy patches should exist and consist the majority of patches. Recently, methods based on DCNN [10] are proposed for document image repair e.g. [11,12] apply DCNN for binarization with models proposed for semantic segmentation. In [11], the experiment demonstrates the proposed model outperforms state of the art methods. However, these two models require annotated data for training and do not provide any interpretation.

In this paper, we proposed a method that fuses DCNN and template matching aim to challenge more severe degradation mainly in foreground without any annotations. As degradation is severe, it is not guaranteed that healthy patterns could be obtained by unsupervised learning like [9]. Leveraging the strong ability of DCNN to generalize, we use an estimated font and simulated degradation to generate data for DCNN training. But unlike [11,12], there is no annotation for our target data and we use simulated data. The trained DCNN provides a useful reference for repairing the real data even though it is only trained on simulated data. In the end, a voting mechanism that makes use of the reference with a template matching method provides interpretation to repair process.

2. Target data

We search target data according to the degradations that we aim to address i.e. the degradations in foreground. The selected target data of our study is from a collection, University of Chicago Laboratory Schools Work Reports 1898-1934[13,14,15], from University of Chicago Library. The documents in the collection are typed by typewriters

Figure 4: Degradation process of a '3'. The operations transform the patch from left to right are dilation, fading and shifting respectively.

(a) Simulated patches with single character

(b) Simulated patches with auxiliary characters besides main character

Figure 5: Examples of simulated degradation patches

and so share the same font and font size. 10 images of pages with different kinds and degrees of degradation are selected from the collection. Each image contains about 1200 characters. Each character occupies about 19 pixels in width and about 40 pixels in width. Figure 2 shows several lines of an image. Figure 3 shows some examples of the degraded characters in the images. The degradation in the document is changeful while the documents are made with the same approach. Among the images, the hard ones are about just right readable by a human. We consider it as an appropriate difficulty for study on this problem as we can still manually annotate documents for evaluation and analyze qualitatively while keeping the problem challenging.

3. Approach

3.1. Overview

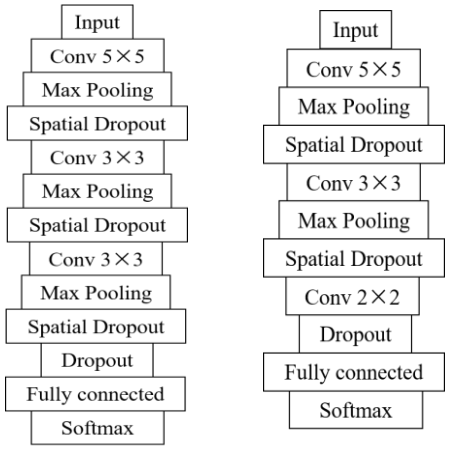
The repair process of a document image mainly consists of 4 steps, namely segmentation, recognition, template matching, and decision making. We assume that the font and font size is estimated for simplicity. This assumption is reasonable to a certain extent as the target data is made with typewriters and so is of a certain font and font size. In the segmentation step, the whole image is first segmented into lines of characters by a thresholding method and then segmented into patches that contain one character by a DCNN. Note that patches contain one character are treated as a part of the image rather than exactly a character. We

Figure 6: A visualization of segmentation process. The first line is original image. White means high belief to be center of a character. The second line is the corresponding response from DCNN. The third line is response after non-maximum suppression. The fourth line is the final result after merge process.

choose such a size for a patch instead of a smaller one because so that patches naturally imply the architecture of characters and are easy to localize. Also, a smaller patch tends to contain only pixels like noise since the images are of low resolution and severely degraded. In recognition step, the patches are classified into the 96 characters with different confidence by the other DCNN. We train both of the DCNNs with simulated degraded patches because on one hand there is no annotation for the target data and on the other hand it is difficult to obtain all the patterns as the frequencies of characters in documents are badly uneven e.g. 'Z' hardly appears. In the template matching step, each prediction of a patch with confidence more than a threshold value is aligned with a healthy template. Finally, in decision making step, based on a voting mechanism the matched templates of a degraded patch decide how to repair that patch. Besides, the estimation of font and font size, as well as the segmentation and recognition of patches, are not necessary to be accurate to obtain the desired repair result, though the quality of repair result is affected by them.

3.2. Degraded patch simulation

As the characters in target data are typed by typewriter, the degradation tends to be consistent among one character. To generate simulated degraded patches, we first generate healthy patches of the estimated font with the estimated font size and then degrade them. The process transforms a healthy patch to a degraded patch contains 4 steps as shown in Figure 4. First, a healthy patch is sampled. Second, the patch dilates for 2 epochs. In each epoch, each of the pixels has a chance to dilate to make its neighboring pixels darker. Third, Gaussian filter shape erosions are randomly applied to a portion of pixels. Finally, a shift operation is performed. The simulated patches for segmentation training and recognition training are different. The patches for segmentation training have two auxiliary characters except for the main character and shift only vertically as shown in figure 5 (a). The patches for recognition training only contain a single character and shift horizontally and vertically as shown in figure 5 (b). Although we design the degradation simulation to mimic the degradation in real condition, the eventual objective is to obtain robust DCNNs rather than to simulate accurate patches.



(a) DCNN for segmentation (b) DCNN for classification

Figure 7: Architectures of DCNNs. The left one is for segmentation and the right one is for recognition.

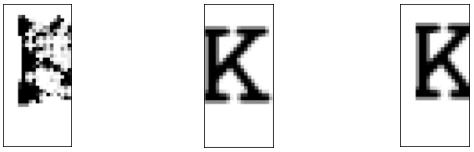


Figure 8: A patch matches template of 'K'. From left to right, the images are a degraded patch, the healthy patch of 'K' and the matched template respectively.

3.3. Segmentation

Before segmented into patches, a document image is segmented into lines. A 3 by 2 times of character width uniform filter was used as a line detector. The lines with pixels with a low response of the filter are considered as lines contain foreground. Neighboring narrow lines are merged until they form a line of enough height. Eventually, images of lines of characters are segmented from the original image.

Character segmentation is performed by a DCNN as a binary classification in a detection way. Given a patch, the DCNN predict whether the patch share center horizontally with a character.

When inference, the responses from DCNN are followed by non-maximum suppression and a merge process, in which the predicted character center merges with each other if the distance between them is less than a threshold as shown in figure 6.

The DCNN we use for segmentation is sequential as shown in figure 7 (a). Spatial dropout [16] and normal

Figure 9: The repair process of a degraded patch. The left image is a degraded patch. The three images in the middle are the matched templates. The right image is the repaired patch.

dropout [17] layers are added to regularize to prompt robustness of the model. In a way, we can think the dropout layers as a kind of fading simulation embedded in the network.

3.4. Recognition

We also use DCNN to classify the patches. For an input patch, beliefs are predicted for each of the 96 classes. Then the predictions with belief more than a threshold are selected as voters to decide repair scheme. The network architecture as showed in figure 7 (b) is quite the same as the one for segmentation.

3.5. Template matching

The selected voters of a patch are matched with a template. Given a voter of a degraded patch, a bunch of patches is generated from the healthy patch shares the same class with the voter by enumerating dilation and shifting. Among all the bunch of patches, the patch that with the minimum similarity score with the degraded patch is assigned as the matched template of the voter. After matching, the dilation in a template is canceled. This process localizes the degraded patches and enables us to cancel dilation. Figure 8 shows an example of a matched template.

$$Cover(A,B) = \begin{cases} \frac{sum\left((1-A)\cdot \min(EDT(B), thres)^2\right)}{sum(1-A)}, & if\ sum(1-A) \neq 0 \\ 0, & if\ sum(1-A) = 0 \end{cases} \quad (1)$$

The similarity score of patch A and patch B is calculated as the mean of the coverage measures of A to B and B to A. To calculate the coverage measure, A and B are first binarized to 0 and 1 for foreground and background respectively with Otsu's method [18]. Then the coverage measure is calculated as formula 1. '1-A' means the reverse of A. EDT stands for Euclidean distance transform, which transforms the binarized patch to a distance map in which each pixel denotes the distance between it and the nearest 1 in the original patch. A threshold is set to tolerate noise in the patch. The coverage measure of A to B essentially measures how well patch B cover patch A. A coverage measure of 0 means the two patches are identical and a high value indicates they are quite different.

	Free Online OCR		Tesseract OCR	
Image name	Original image accuracy	Repaired image accuracy	Original image accuracy	Repaired image accuracy
012-8	27.01	35.31	85.79	85.63
029-4	45.98	43.74	95.78	89.83
033-3	38.04	41.51	88.62	85.39
033-4	45.41	43.59	95.46	92.60
mean	39.11	40.99	91.41	88.36

Table 1: Character level OCR accuracy of original and repaired images

3.6. Repair

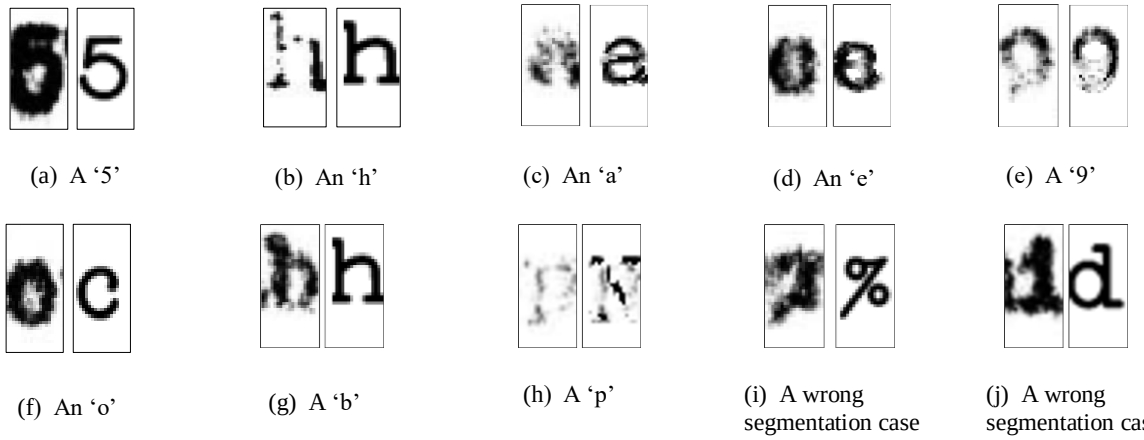
The repair scheme of a patch is determined by the matched templates of that patch. The matched templates vote pixel by pixel. If all the pixels of a position in all the matched templates are similar, the pixel of that position in the degraded patch would be replaced by the mean grey scale value of that position of the matched templates.

Figure 9 shows an example of a repair process. The degraded patch in figure 9 is classified as 'O','G','C' with high beliefs. The matched 'O','G','C' templates all contain a 'C' shape ellipse curve and are not dilated, so the corresponding part in the degraded patch is replaced with that shape and the dilation is canceled while the positions that the templates are different on remain unchanged. This process is inspired by the fact that neural networks predict based on features. Due to the complexity of the features leverage by neural networks, they provide low interpretation. However, it is likely that the reason for a network to predict multiple class with high confidence is that the predicted classes share certain features that exist in input. This voting mechanism tries to extract the shared features of the predictions with high belief, which is more trustworthy and is the proof the network relies on for prediction. By this mechanism, we provide security and interpretation in a way.

4. Experimental Results and Discussions

To emphasize the effect of the proposed method, we only evaluate the four hardest images in the target data. A standard typewriter font is picked as the estimated font. The patch size is set to 19 pixels in width and 40 pixels in height.

Figure 10 shows some success and failure cases in the experiment. From the qualitative results we can see that the segmentation and recognition networks are robust to dilation, diffusion and fading to some degree like in Figure 10 (a), (b). The voting mechanism helps to repair partially or reject repair like in Figure 10 (f) when the recognition



(a) A '5' (b) An 'h' (c) An 'a' (d) An 'e' (e) A '9'

(f) An 'o' (g) A 'b' (h) A 'p' (i) A wrong segmentation case (j) A wrong segmentation case

Figure 10: Some patches in the original images(left) and corresponding repaired patches(right). The first row are successful cases and the second row are failed cases.

is with much uncertainty and sometimes lead to quite good results when the top-1 prediction is incorrect or the correct prediction doesn't even appear in top-3 predictions like in Figure 10 (c), (d). From the failure cases, we can see that wrong segmentation or recognition happens from time to time like the second row in Figure (10). It may be a wrong way to do segmentation in a detection way that it doesn't really segment the patches and so a patch may contain a small part of the other patch, which leads to wrong recognition like in Figure 10 (j). Besides, as we set a fixed window size of a patch the window sometimes does not contain a whole patch when the patch is wide and the center is shifted. Therefore, instead of finding the centers of patches, to find the boundaries may make more sense in terms of segmentation. Errors of recognition happen more frequently when the patches are faded like in Figure 10 (g), (h). This may be caused by the fact that the simulation of fading is far away from reality. The recognition network now classifies a patch into 96 classes i.e. to give each that sum up to one to each of the class. This doesn't match our repair method as we don't actually want to know the top-1 result but which classes are with higher beliefs. Therefore, to do binary classification for each of the class is more reasonable then do multi-class classification.

Table 1 shows quantitative results as the effects of our method on OCR systems. The OCR system we use includes a free online OCR [19] and Tesseract OCR [20], which is a mainstream open source OCR system that keeps developing. The OCR results are evaluated by the ISRI analytic tools for OCR Evaluation [21] at character level. While the proposed method improves results slightly for the free online OCR, it decreases the performance of Tesseract OCR. For both of the OCR systems, the proposed method obscure recognition in some cases. It is not surprising that our method doesn't work well as we only consider the OCR system as a way to evaluate rather than designed to improve OCR system. To make the proposed method more compatible with OCR systems, the dedicated design is required. The degradation types, mainly diffusion, in the first image are considered more so the method performs much better for the first image. The other images contain more severe dilation and fading.

In words, despite that the performance is poor now, the qualitative result shows that the method somewhat makes sense. The project now is nearly a prototype. As the target data is challenging, much adjustment is required.

5. Conclusions and Future Work

We presented a novel approach to document image repair fusing DCNNs and template matching that can repair severely degraded document image without annotated training data. Through fusion, the high performance of DCNNs and the interpretability of template matching are combined.

The assumption that the font and font size is manually estimated impedes the generalizability of the method. Automatic font and font size estimation, or in other words to learn patterns from input similar to [9] but also handle the cases that it is hard to totally learn from inputs, is required to make the method elegant as well as to improve performance. It is considerable to mix the real data and the simulated data for training and to estimate healthy patterns since though the patterns in real data may not be complete, the patterns frequently appear in real data make much sense. Currently, segmentation and recognition are separated in the system. In consideration of the success of multitask learning [21], we believe that to integrate them into one DCNN improves performance e.g. to do them as a

semantic segmentation task or recognize in a sequential way. To form a complete system for the realistic application, a language model is required to be plugged in. A language model is expected to improve the performance and is necessary in some cases e.g. the cases in which faded characters and bleed-through characters both appear and may not be distinguishable. In the study, we are aware that different kinds of degradation call for widely different repair process e.g. dilation and fading, so to classify degradation and to repair them in specialty for different degradations is promising.

References

[1] Sankar, K.P., Ambati, V., Pratha, L. and Jawahar, C.V., 2006, February. Digitizing a million books: Challenges for document analysis. In International Workshop on Document Analysis Systems (pp. 425-436). Springer, Berlin, Heidelberg.

[2] Shi, Z. and Govindaraju, V., 2004, August. Historical document image enhancement using background light intensity normalization. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 1, pp. 473-476). IEEE.

[3] Shukla, S., Sonawane, A., Topale, V. and Tiwari, P., 2014. Improving degraded document images using binarization technique. International Journal of Scientific & Technology Research, 3(5), pp.333-338.

[4] Howe, N.R., 2013. Document binarization with automatic parameter tuning. International Journal on Document Analysis and Recognition (IJDAR), 16(3), pp.247-258.

[5] Ioannis Pratikakis, Konstantinos Zagoris, Panagiotis Kaddas and Basilis Gatos. ICFHR 2018 Competition on Handwritten Document Image Binarization (H-DIBCO 2018). 16th International Conference on Frontiers in Handwriting Recognition (ICFHR'18).

[6] Tu, H.Y., Hsia, C.H. and Giang, H.T.H., 2016. An efficient adaptive image enhancement method in wavelet domain for handwritten document. Journal of Applied Science and Engineering, 19(3), pp.357-370.

[7] Moghaddam, R.F. and Cheriet, M., 2010. A variational approach to degraded document enhancement. IEEE transactions on pattern analysis and machine intelligence, 32(8), pp.1347-1361.

[8] Babu, N., Preethi, N.G. and Shylaja, S.S., 2008, December. Degraded document image enhancement using hybrid thresholding and mathematical morphology. In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing (pp. 701-705). IEEE.

[9] Banerjee, J., Namboodiri, A.M. and Jawahar, C.V., 2009, June. Contextual restoration of severely degraded document images. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 517-524). IEEE.

[10] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[11] Meng, G., Yuan, K., Wu, Y., Xiang, S. and Pan, C., 2017, November. Deep networks for degraded document image binarization through pyramid reconstruction. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 727-732). IEEE.

[12] Tensmeyer, C. and Martinez, T., 2017, November. Document image binarization with fully convolutional neural networks. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 99-104). IEEE.

[13] University of Chicago. Laboratory Schools. Work Reports, [Box 2, Folder 33], Special Collections Research Center, University of Chicago Library.

[14] University of Chicago. Laboratory Schools. Work Reports, [Box 2, Folder 29], Special Collections Research Center, University of Chicago Library.

[15] University of Chicago. Laboratory Schools. Work Reports, [Box 2, Folder 12], Special Collections Research Center, University of Chicago Library.

[16] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), pp.1929-1958.

[17] Tompson, J., Goroshin, R., Jain, A., LeCun, Y. and Bregler, C., 2015. Efficient object localization using convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 648-656).

[18] Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics, 9(1), pp.62-66.

[19] https://www.onlineocr.net/

[20] Smith, R., 2007, September. An overview of the Tesseract OCR engine. In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) (Vol. 2, pp. 629-633). IEEE.

[21] Rice, S.V. and Nartker, T.A., 1996. The ISRI analytic tools for OCR evaluation. UNLV/Information Science Research Institute, TR-96-02.

[22] Ruder, S., 2017. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.