Machine Perception DS/NC/ESD 863 Denoising Dirty Documents

Project Review

April 18, 2017

List of Figures

1	Sample Images	2
2	Image 3 and its Histogram	4
3	Image 5 and its Histogram	4
4	Histogram Suppression a)	5
5	Histogram Suppression b)	5
6	Histogram Suppression c)	5
7	Fixed Thresholding with Multiple Threshold Values	6
8	Fixed Thresholding a)	7
9	Fixed Thresholding b)	7
10	Fixed Thresholding c)	7
11	Fixed Thresholding on Coffee Stained Image	8
12	Adaptive Thresholding with Multiple Cleaned Images	9
13	Adaptive Thresholding a)	10
14	Adaptive Thresholding b)	10
15	Canny Edge Detection a)	11
16	Canny Edge Detection b)	11
17	Step by Step Transformations	12
18	Morphological Operation a)	12
19	Morphological Operation b)	13

List of Tables

1 Goal Statement Description

Optical Character Recognition (OCR) is the process of getting typed or handwritten documents into a digitized format. The motivation of converting to a digitized format is to ensure security, accessibility, edit-ability and ease of searching and sharing. Also, digital documents don't get dirty and cannot be ruined by coffee stains. [2]

Unfortunately, a lot of documents eager for digitization are being held back. Coffee stains, faded sun spots, dog-eared pages, and lot of wrinkles are keeping some printed documents offline and in the past.

Given a dataset of images of scanned text (synthetic images) that are "noisy" with stains and wrinkles, we propose to clean up the noise and help with the digitization process.

2 Dataset

Kaggle provided a dataset which consists of two sets of images - train and test. These images contain various styles of text, to which synthetic noise has been added to simulate real-world, messy documents. The dirty images contain stains as well as creased paper. The training set also includes the cleaned up images of those found in the test file (train_cleaned) [2]. By clean, we mean black letters on a white background. Below are two sample images:

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit is located below the form or they can be printed directly on the fa separate sheet is much better from the point of view of the quabut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, we be taken into account: The best way to print these light rectangles.

(a) Original Image

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit socated below the form or they can be printed directly on the form a separate sheet is much better from the point of view of the que but requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles.

(b) Cleaned Image

Figure 1: Sample Images

Kaggle calculates the score based on the root-mean-squared-error (RMSE) value between each pixels of the generated output and the actual cleaned image.

3 Literature Review

The project is one of *Kaggle's* old competition. *Kaggle* has a section called Kernels & Discussions for each dataset which contains snippets of codes, discussions between participants, ideas for future expansion.

We came across a blog [1], which contained the methods used by the author to solve this. He has used a variety of techniques to solve this. We proposed our methods based on our understanding of the problem and the approach taken by him and the other users.

4 Methods

Taking the proposed plan of action forward, we have implemented 4 methods mentioned in our previous project report. We have identified the methods that have worked well and for which images and the challenges we have faced.

4.1 Histogram Suppression

To understand the data-set, we plotted the images alongside their histograms. It gives us a clearer understanding of the noise in the images.

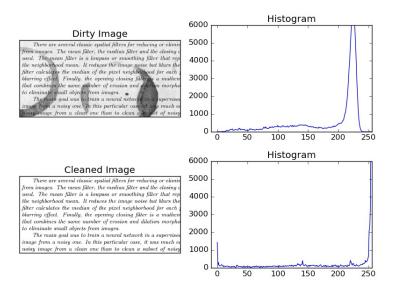


Figure 2: Image 3 and its Histogram

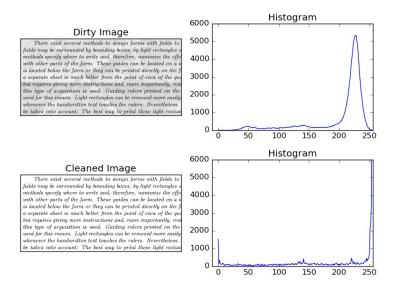


Figure 3: Image 5 and its Histogram

As you can see from the plots, most of the pixels are concentrated near the 200's there are some scattered around the range (50 to 150). This pattern arises in all of the images, some more denser than others. This leads us to conclude that the intensity of the characters are in this range (Note, noise could also be found in this region).

We propose an algorithm that changes all pixel values to 255 (white) leaving only the ones in the range 50 to 150 unchanged. The results are displayed below.

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectange the surrounded by bounding boxes, by light rectange the surrounder of the form. These guides can sheet of paper that is located below the form or they can be form. The use of guides on a separate sheet is much better from quality of the scanned image, but requires giving more instruction restricts its use to tasks where this type of acquisition is used. The form are more commonly used for this reason. Light remore easily with filters than dark lines whenever the handwritte Nevertheless, other practical issues must be taken into account

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectang. These methods specify where to write and, therefore, minimi overlapping with other parts of the form. These guides can sheet of paper that is located below the form or they can be form. The use of guides on a separate sheet is much better from quality of the scanned image, but requires giving more instruction restricts its use to tasks where this type of acquisition is used. The form lare more commonly used for this reason. Light remore easily with filters than dark lines whenever the handwritts Nevertheless, other practical issues must be taken into account

(a) Original Image 1

(b) Cleaned Image 1

Figure 4: Histogram Suppression a)

There are several classic spatial filters for reducing or elimin from images. The mean filter, the median filter and the closing of used. The mean filter is a lowpass or smoothing filter that repl the neighborhood mean. It reduces the image noise but blurs the filter calculates the median of the pixel neighborhood for each p blurring effect. Finally, the opening closing filter is a mathem that combines the same number of erosion and dilation morphot to eliminate small objects from images.

to eliminate small objects from images.

The main goal was to train a neural network in a supervised image from a noisy one. In this particular case, it was much ea noisy image from a clean one than to clean a subset of noisy

There are several classic spatial filters for reducing or elimin from images. The mean filter, the median filter and the closing of used. The mean filter is a lowpass or smoothing filter that repl the neighborhood mean. It reduces the image noise but blurs the filter calculates the median of the pixel neighborhood for each p blurring effect. Finally, the opening closing filter is a mathemathat combines the same number of crossion and dilation morphol to eliminate small objects from images.

The main goal was to train a neural network in a supervised image from a neisy one. In this particular case, it was much ea noisy image from a clean one than to clean a subset of noisy

(a) Original Image 2

(b) Cleaned Image 2

Figure 5: Histogram Suppression b)

There exist several methods to design to be filled in. For instance, fields may be bounding boxes, by light rectangles or b. These methods specify where to write and mize the effect of skew and overlapping w. the form. These guides can be located on a paper that is located below the form or the directly on the form. The use of guides on is much better from the point of view of t scanned image, but requires giving more more importantly, restricts its use to task

(a) Original Image 3

There exist several efficients to design to be filled in. For instance, fields may hounding body, by light rectanges on both the methods specify where to write and mise the effector saws and overlapping withe forth. These guides can be located on a paper has a located below the form or the checky on the form. The use of guides on is much better from the point of view of it seemed image, but it is the forthesis.

(b) Cleaned Image 3

Figure 6: Histogram Suppression c)

The RMSE value turned out to be 239.53297.

4.2 Fixed Thresholding

In order to separate the writing from the background noise, we first tried out the fixed thresholding method. We recognised that the writing was significantly darker than

the background which contained noise in the form of creases or dog eared pages. The biggest challenge here was the trial and error involved in deciding on a good value to threshold the image on. Below are the different potential values that gave us a reasonable output.

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit is located below the form or they can be printed directly on the feat a separate sheet is much better from the point of view of the quebut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles.

(a) Original Image 1

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit located below the form or they can be printed directly on the fixed a separate sheet is much better from the point of view of the quabut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, a be taken into account: The best way to print these light rectangles.

(c) Cleaned Image with Threshold 160

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit located below the form or they can be printed directly on the fix a separate sheet is much better from the point of view of the que but requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless. be taken into account: The best way to print these light rectang

(e) Cleaned Image with Threshold 170

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a si located below the form or they can be printed directly on the fc a separate sheet is much better from the point of view of the quabut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles.

(b) Cleaned Image with Threshold 155

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to urite and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit located below the form or they can be printed directly on the fix a separate sheet is much better from the point of view of the que but requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless.

(d) Cleaned Image with Threshold 165

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit ocated below the form or they can be printed directly on the fix a separate sheet is much better from the point of view of the que but requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless.

(f) Cleaned Image with Threshold 175

Figure 7: Fixed Thresholding with Multiple Threshold Values

We decided that out of these values, a threshold value of 165 gave us the best result for the most images. Below are the dirty and cleaned image pairs after fixed thresholding with a threshold value of 165.

ret, thresh1 = cv2.threshold(img,165,255,cv2.THRESH_BINARY)

There are several classic spatial filters for reducing or elimin from images. The mean filter, the median filter and the closing o used. The mean filter is a lowpass or smoothing filter that repeture neighborhood mean. It reduces the image noise but blurs the filter calculates the median of the pixel neighborhood for each plurring effect. Finally, the opening closing filter is a mathem that combines the same number of erosion and dilation morpho to eliminate small objects from images.

The main goal was to train a neural network in a supervised image from a noisy one. In this particular case, it was much ea noisy image from a clean one than to clean a subset of noisy

(a) Original Image 2

There are several classic spatial filters for reducing or elimin from images. The mean filter, the median filter and the closing o used. The mean filter is a lowpass or smoothing filter that repethe neighborhood mean. It reduces the image noise but blurs the filter calculates the median of the pixel neighborhood for each 1 blurring effect. Finally, the opening closing filter is a mathem that combines the same number of erosion and dilation morpho to eliminate small objects from images.

The main goal was to train a neural network in a supervisea image from a noisy one. In this particular case, it was much ea noisy image from a clean one than to clean a subset of noisy

(b) Cleaned Image with Threshold 165

Figure 8: Fixed Thresholding a)

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles a methods specify where to write and, therefore, minimize the effectivith other parts of the form. These guides can be located on a sis located below the form or they can be printed directly on the fast separate sheet is much better from the point of view of the quebut requires giving more instructions and, more importantly, resistively of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles.

(a) Original Image 3

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a, is located below the form or they can be printed directly on the fix a separate sheet is much better from the point of view of the quabut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever, the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles.

(b) Cleaned Image with Threshold 165

Figure 9: Fixed Thresholding b)

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit located below the form or they can be printed directly on the far separate sheet is much better from the point of view of the quabut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles.

(a) Original Image 4

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit located below the form or they can be printed directly on the fix a separate sheet is much better from the point of view of the quabut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles.

(b) Cleaned Image with Threshold 165

Figure 10: Fixed Thresholding c)

The problem with this method that it worked only when the noise was in the form of creases or dog eared pages i.e., when the background was much lighter that the writing. However, when the noise is in the form of coffee stains, fixed thresholding didn't work as seen in the image below. This is because the noise was not much lighter than the writing itself. If we increase the threshold value to one that is high enough to remove the coffee stains, then it may remove the dark writing along with the stains.

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sits located below the form or they can be printed directly on the fica separate sheet is much better from the point of view of the quabut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles.

There exist several methers design forms with fields to fields may be surrounded by counding boxes, by light rectangles o methods specify with the write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit located below the form or they can be printed directly on the fa a separate sheet is muchibetter from the point of view of the quabut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the located by the statement of the point of the poi

(a) Original Image 5

(b) Cleaned Image with Threshold 165

Figure 11: Fixed Thresholding on Coffee Stained Image

The RMSE value is 220.00782.

4.3 Adaptive Thresholding

For images which contain coffee stains, we proposed to use adaptive thresholding. This is because although the coffee stains are darker than creases, they are definitely not as dark as the writing. Since adaptive thresholding looks at the neighbourhood of a pixel to do the thresholding, the neighbourhood of the coffee stains is definitely lighter than if we look at the neighbourhood of the writing. Below are the different values for 'box size' and 'constant c' that we tried with both Gaussian and Mean thresholding.

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit ocated below the form or they can be printed directly on the fix a separate sheet is much better from the point of view of the quabut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles.

(a) Original Image 1



There exist several method to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles of methods specify where to write and, therefore, minimize the effectivith other parts of the form. These guides can be located on a sis located below the form or they can be printed directly on the first separate sheet is much better from the point of view of the guidest can be located on a six but requires guing more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this greater tight rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be lakentified account. The best very to print these light rectangles.

- (b) Gaussian Box Size 9 and Constant 2

 There end, several method to design forms with fields to fields may be surpunded by bounding boxes; by lightly ctangles of methods specify whereto write and; therefore minimize the effect with other, parts of the form. These guides can be located on a sis located below the form, or they can be printed directly on the first a separate sheet is much better from the point of view of the quebut requires guing more instructions and, more importantly, rest thus type of acquisition is, used "Guiding rulers fromed on the trust of others by some on the surface of the second on the second of the
- (c) Mean Box Size 9 and Constant 2



(d) Gaussian Box Size 11 and Constant 2

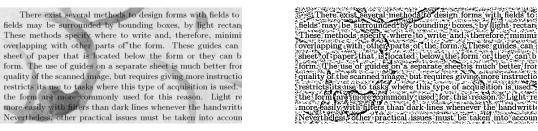
takentinto account: The best way to print these light rectan

(e) Mean Box Size 11 and Constant 2

Figure 12: Adaptive Thresholding with Multiple Cleaned Images

We decided that out of these values, a combination of Box Value of 11 and Constant value of 2 with Gaussian Filtering gave the best output for most images. Below are the dirty and cleaned image pairs after adaptive thresholding.

 $thresh2 = cv2.adaptiveThreshold(img, 255, 1, cv2.THRESH_BINARY, 11, 2)$



(a) Original Image 2

(b) Cleaned Image

Figure 13: Adaptive Thresholding a)

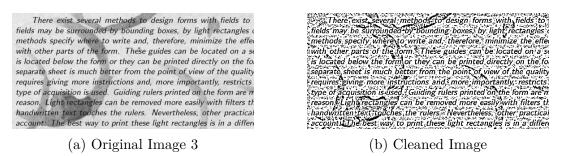


Figure 14: Adaptive Thresholding b)

However, from the above images we see that adaptive thresholding has left a pattern of specks in the place of the coffee stains which further need to be cleaned. Calculated RMSE value turned out to be 212.28749.

4.4 Canny Edge Detection and Morphology

Although the stains are cleaned up to a large extent we still see a bunch of specks which need to be addressed. As proposed, we decide to tackle this problem by first applying Canny Edge Detection on the images. We did trial and error on the min and max value for the Canny Edge Detector. Below are the different potential value pairs that we thought gave us the best result.

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sit located below the form or they can be printed directly on the fa separate sheet is much better from the point of view of the quabut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, we be taken into account: The best way to print these light rectangles.

(a) Original Image 1

There exist-several methods to design forms with fields to fleds may be surrounded by formaling tense, by light restangles westly orders to write each, therefore, minimize the off with other parts of the form. These guides can be bonded on a is localed below the form or they can be grinted directly on the gas approach cheat is much tatter from the guide, of view of the guidences and more instructions and more instructions that the first type of aggustifies is used. Outday release printed on the used for phistication. Light rectangles can be removed more easily whenever this to structure to the total state of the there into any the total state or the total state of the t

(c) Minimum 100 Maximum 400

There exist several methods to Resign forms with fields to fields may to burrounifeld by brivaling briss, by light restangles westhods specify where to write and, therefore, ministical the affer with other parts of the form. These guides can be bouled on a is located below the form or they can be pristed directly of the guide school is neach fatter from the paint, of view of the guide reprine gising more instructions and more insportantly, resthis type of against from it used for phiatronom. Light rectangles can be removed more easily when the paint of the hundridge that the best way to privat these light rectangles in the polary into account. The best way to privat these light rectangles in the paint.

(b) Minimum 100 Maximum 300

There exist several matheds to design forms with fields to fields may be surrounded by tomading boxes, by light yell angles o walleds mostly where to write each, therefore, aministe the offer with other parts of the form. These guides can be bonded on as is located below the form or they can be privated directly on the fit expends should in much batter from the gold of view of the quebut replies giving more much fatter from the point of view of the quebut replies of acquisition is used. Guiding rules privated on the used for their cools. Light rectangles can be removed more enough the private of the particular of the particular of the cools.

(d) Minimum 100 Maximum 500

Figure 15: Canny Edge Detection a)

We decided that for a minVal of 100 and maxVal of 500, the speckled part was getting cleaned up for most of the images.

edges = cv2. Canny (img, 100, 200)

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectange. These methods specify where to write and, therefore, minimi overlapping with other parts of the form. These guides can sheet of paper that is located below the form or they can be form. The use of guides on a separate sheet is much better fror quality of the scanned image, but requires giving more instruction restricts its use to tasks where this type of acquisition is used, the form are more commonly used for this reason. Light remore easily with fifters than dark lines whenever the handwritts. Nevertheless, other practical issues must be taken into account

(a) Original Image 2

There exists several methods to design forms with fields to fields may be surrounced by bounding bases, by light rectan I been methods excellent where to write and, therefore, minimit overlapping with other parts of the form. These guides can sheet of paper that is located below the form or they can be form. The use of guides in a separate sheet, is much better from quality of the seamed image, but requires giving more instruction restricts its use to take where this type of equipition is used the form are not recommonly used for this reason. Light remove deathy titl theirs than dark lines whenever the handwrite Meyerfuldess? offer proceeds a season much sense much be taken into account

(b) Cleaned Image

Figure 16: Canny Edge Detection b)

What we notice is that two edges are formed around the writing, like a stencil. While the coffee stains have only a single edge. We attempted we use morphological techniques of erosion and dilation in order to get rid of the stain edges completely. We tried dilating the image once, in order to make the writing much thicker than the edges (as the writing had two edges that would expand) and then eroding the image in order to get stains while keeping the writing intact due to the thickness difference. Below is an example of the step by step process followed.

```
kernel = np.ones((3,3),np.uint8)

dilation = cv2.dilate(edges, kernel, iterations = 1)

erosion = cv2.erode(dilation, kernel, iterations = 1)

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles of methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sis is located below the form or they can be printed directly on the fix a separate sheet is much better from the point of view of the quebut requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this russion. Eight rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles (a) Original Image 1

(b) Canny Edge Detection
```

Figure 17: Step by Step Transformations

Below are the dirty and cleaned image pairs after dilation and erosion are done.

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectang These methods specify where to write and, therefore, minimic overlapping with other parts of the form. These guides can sheet of paper that is located below the form or they can be form. The use of guides on a separate sheet is much better fror quality of the scanned image, but requires giving more instruction restricts its use to tasks where this type of acquisition is used, the form are more commonly used for this reason. Light remore easily with filters than dark lines whenever the handwritte Nevertheless, other practical issues must be taken into account

(c) Dilation

(a) Original Image 2

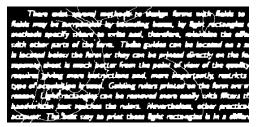
There acids are real methods to testing forms with fields to being some, by fight restant these sections appelly where to write and, therefore, minimizes sections appelly where to write and, therefore, minimizes of the form. These galdes can sheet of paper that is boused below the form or they can be orn. The use of galdes on a superate sheet in some better for pulley of the seas sections, but require giving norm insertucion writings in tag to joint where this type of sequilibrium is used. The form of more portenously used for this resistent, Light, respectively with higher than durk inconsiderate it is hundwrite.

(d) Morphological Operation

(b) Cleaned Image

Figure 18: Morphological Operation a)

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles a methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a sail located below the form or they can be printed directly on the for separate sheet is much better from the point of view of the quality requires giving more instructions and, more importantly, restricts type of acquisition is used. Guiding rulers printed on the form are meason. Light rectangles can be removed more easily with filters the handwritten text touches the rulers. Nevertheless, other practical account: The best way to print these light rectangles is in a different methods.



(a) Original Image 3

(b) Cleaned Image

Figure 19: Morphological Operation b)

However we see that the writing isn't clear in the cleaned up image after the final erosion is done. Hence, we decided to avoid doing the morphological operations.

RMSE value for dilation is 196.74589.

RMSE value for erotion is 218.05208.

Methods/Score	RMSE
Histogram Suppression	239.53297
Fixed Thresholding	220.00782
Adaptive Thresholding	212.28749
Canny Edge (Dilation)	196.74589
Canny Edge (Erotion)	218.05208

Table 1: Table with methods and their RMSE scores

5 Future Work

Until now, we focused on using image processing techniques to "denoise" the dirty documents. We propose to take a few machine learning focused approaches to see if it gives us an improved denoising ability. One approach we thought of is the prediction of the brightness of a pixel (black for writing and white for background) based on the 9 surrounding pixels. We can use this information on different supervised learning models and see if this gives us good results. We also believe that we can integrate this local pixel information with known properties about the structure of the document like line or word spacing to improve our models.

References

- [1] Colin blog. http://tinyurl.com/gnptby6. Accessed: 2017-04-18.
- [2] Kaggle denoising dirty documents. http://tinyurl.com/z4ukatx. Accessed: 2017-04-18.