Sparse Ensemble Learning with Truncated Convolutional Autoencoders for Cleaning Stained Documents

Anant Khandelwal [1]

[1] *Electrical Engineering Department, Indian Institute of Technology Delhi, Shaheed Jeet Singh Marg, Hauz Khas, New Delhi, Delhi 110016, India*
[1] *jtm162085@dbst.iitd.ac.in*

Abstract

This paper mainly focus on how to extract clean text from the stained document. It may happen sometimes that due to stains it becomes very difficult to understand the documents and from the previous work it has been seen that one particular modelling technique either through Image processing or Machine learning which alone can't perform for all the cases in general. As we all know ensemble techniques combine many of the modelling techniques and result in much reduced error that would not be possible by just having single model. But the features used for different models should be sparse or non-overlapping enough to guarantee the independence of each of the modelling techniques. XGBoost is one such ensemble technique in comparison to gradient boosting machines which are very slow due to this it's not possible to combine more than three models with reasonable execution time. This work mainly focus on combining the truncated convolutional Autoencoders with sparsity take into account to that of machine learning and Image processing models using XGBoost such that the whole model results in much reduced error as compared to single modelling techniques. Experimentation's are carried out on the public dataset NoisyOffice published on UCI machine learning repository, this dataset contains training, validation and test dataset with variety of noisy greyscale images some with ink spots, coffee spots and creased documents. Evaluation metric is taken to be RMSE(Reduced Mean Squared Error) to show the performance improvement on the variety of images which are corrupted badly.

Keywords: Machine Learning, Deep Learning, Image Morphology, k-means clustering, convolutional Autoencoders, XGBoost, Sliding Window, Random Forest, Gradient Boosting Machines

1. Introduction

Written text is basically a basic form of human communication in documents. Noisy documents seem hectic to read at times also there are important documents which may be mistakenly corrupted by things like coffee stains, dirt, ink spots. In this work we explore the recent advancements in computer vision and machine learning to arrive at the cleaned form of document which exactly retain the text the original while be able to remove dirt completely. Also sometimes we have to further do processing with the handwritten text such as Neural machine language translation, language modelling and for that it is necessary to remove all the stained parts and leave just the writing so that further tasks become very easy. There are useful information on images like car plate number detection, extracting news from very old newspaper, there are complex background containing text which got lost in the background, as text represents the semantic information about image it would be much useful to extract the clean text from image. From all this scenarios scene[37,38,39,40] containing text is the most difficult task as it contains complex background sometimes darker than writing that it would be impossible to recognize text from the image processing techniques or with just classification. Most of the work present in the current literature are in the form of text extraction based techniques using image processing like using Sober filters, Gabor filters, Discrete Cosine transform, Discrete wavelet transform and then leave to just one classifier like SVM[35] or k-means Clustering[27] to form cluster of text which we will show itself is not a complete approach because with the current advancements in the area of machine learning we can actually learn the features of dirt instead of using image processing techniques to remove them the obvious advantage we will show in the further sections in the paper that image processing by itself not be able to completely remove the stains as they are currently rely on either edge based features, or component based features. Edge based features[22-35] based on finding the edges all around text and stains as stains contain only single pixel wide edge whereas text consists of two edges running parallel to each other so applying image morphology techniques dilation make the text thicker as compared to stains, leave out most of the writing. On the other hand component based feature[36-41] are based on separating the text from the stains using some of the clustering techniques to identify the components of text and stain using many of the techniques like Sobel filters, Gabor filters, Discrete Cosine Transform, Discrete Wavelet Transform[25]-[33] and then using these features to feed in neural network or SVM, but both of these depends on the complex background whether the text is identifiable and these techniques are slow also, another technique using K-means clustering by thresholding which we will explored in this work but it has been seen that even that not been able to recover text from complex background satisfactorily. The rest of the sections in this paper are as follows section 2. detailed the related work in this field. Section 3. the Proposed approach through combining many techniques using sparse ensembling these are 1) Linear Regression 2) K-means Clustering Threshold 3) Image Morphology 4) Sliding Window 5) Truncated Sparse Convolutional Autoencoder. Finally section 4. shows results & finally conclusion.

2. Related Work

Recent developments in the area of computer vision and machine learning lot of efforts has been made by researchers towards this end the probabilistic graphic modelling[11,12] approaches and variational EM. For example document denoising by Markov Random Fields[1,10] for learning the statistical distribution of text around whole of the text but the problem with this approach the text may be placed at random position in the images which have complex background and those are meant for something describing the image that can be dealt with sparse coding approaches[2]. Sparse coding approaches like Gabor Filters[13] does the work like duplicating the visual structure at arbitrary location in the document that can eventually be helpful in detecting the text but comes to the problem of cleaning the whole document that doesn't serve either. Obviously there are other approaches too which can model the exact representation of patterns or text to clean the document but but the problem is that we can,t generalize them as they need to accurately consider each of the scenarios like [3]-[8] uses the hidden variables to model the pattern around whole the document. Lot of researcher also tried to use the neural network and CNN[15,16] to clean out the text from the arbitrary background. Wei and lin[51] proposed the robust approach for text detection from video frames it uses 1) identify the locations of text using gradient and K-means clustering 2) separate text from the noisy part using thresholding and SVM, but results have lots of false positives. Xu & Su[52] performs a hierarchical text detection and boosted CNN filtering to identify the location of text but again the complex background case does not show remarkable improvement. To consider the case of natural scene images[46] symmetry based methods are used to detect lines in the text but this doesn't perform well in case of low illuminated text. To mitigate this and considering low illumination [53] proposed a method by locally identifying text based on segmentation but this fails to detect the characters having curvy lines. There is another classification based method[29]-[30] which considers the low and high contrast regions of text by considering a proper threshold value but the problem is threshold value changes as the region changes. There is another work to take into account the nearby pixels using sliding window[43] it takes the features using sliding window and apply SVM to classify text but the method fails to detect non-horizontally aligned text. Another method to classify text[42] is first identifying edges using wavelet transform and K-means clustering to identify text fails either to shown any remarkable improvement. The problem to localize text accurately in case of natural scene images [36] has been done by conditional random fields and energy minimization but the method is complex. Another work with the use of Sobel edge detector and block classification[28] failed to detect the text when text and background have similar intensity. There is lots of work to do some preprocessing to identify locations of text and then classification[13,21,44,22,23,20] but unfortunately this approach doesn't work either due to more false positives or doesn't generalize the case. Another method of generative modelling approaches proposed by William and Titsias[8], Jojic and Frey[6,9,10,17,18,19] but they have the problem of identifying text only in the case of static background only i.e. it doesn't provide a mechanism to model noisy patterns in the document which make it worse for our case. Another approach using variational EM[12] failed to perform when the text is more and more irregular. Yet another method to classify text and non-text regions is proposed[24] it uses dual tree based DWT for edge detection and then classify the text but it takes only the edge based feature for classification that's why the overall performance is not good. This completes the the literature review from the review it has been identified that although there has been much work has been done to extract the text from noisy background but there is still going a search for more robust algorithm which generalize the case or we can there is search for master algorithm to clean document whatever the background and irrespective of how much noisy is the document.

3. Proposed method

Proposed method shown in figure 1. detailing the model we are going to combine in further sections. These models are chosen to maintain the sparsity of model and computational complexity which will be discussed in next few sections.

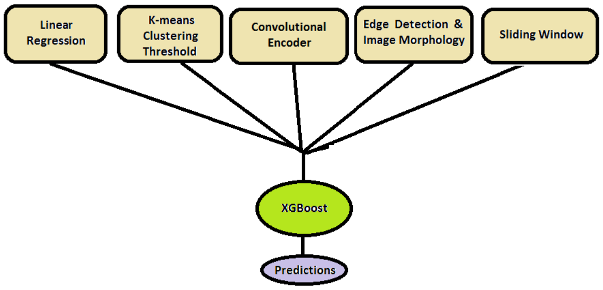
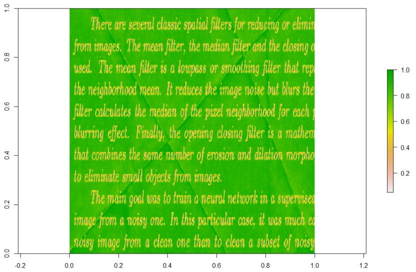


Fig. 1. Proposed System

3.1. *Linear Regression*

Since we have training dataset comprising dirty images and clean images and it is known that images are stored as matrices with row and column identifier denoting the brightness of the pixel at that location. Lets take the image as being three dimensional surface comprising row identifier along x -axis column identifier along y-axis and the brightness of the pixel along z-axis. As we have both dirty and clean images in the dataset, so it is just transforming one surface to another as we have the pixel brightness of both dirty and clean images. Further if we represent the image in the vectorized form so that let's say x vector contains the pixel brightness of a dirty image and vector y represent the brightness corresponds to cleaned one. We can do this easily as from the figure 2. the pixel brightness values remain between $[0, 1]$, with 0 being the brightness value for black and 1 being white.



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 rep
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 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

Fig. 2. Heat Map for Pixel Brightness

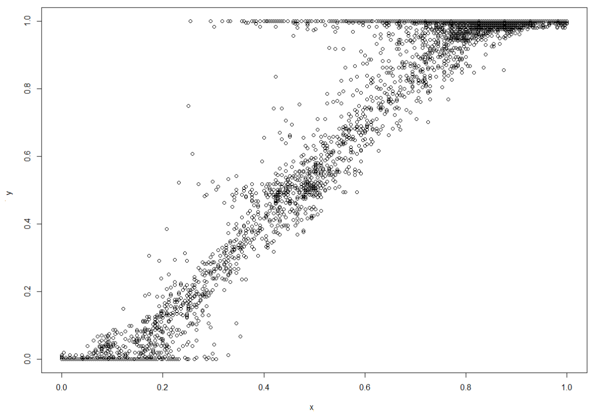
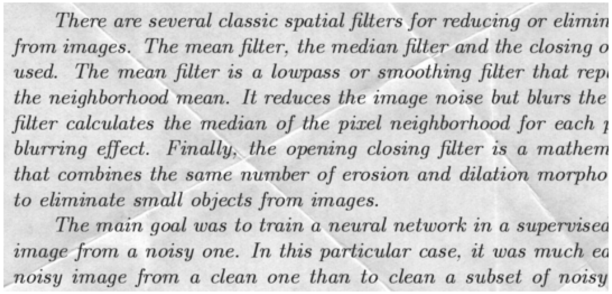


Fig. 3. Relation between x and y

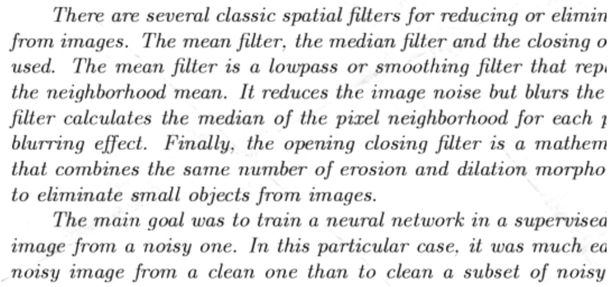
From figure 3. it has been seen that for the sample image the relation between x and y is mostly linear except at the extremes, so it is justifiable to use truncated linear regression. We have used the range for x to be between [0.30.8]. The sample image cleaned is shown in the figure 5, compared to dirty image in figure 4. Considering this model mainly does the work of contrast and brightness correction. The RMSE for the training dataset is calculated as 0.157, this reduces the RMSE from 0.157 to 0.0778.



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 rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel neighborhood for each
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

Fig. 4. Original dirty image

Predicted image for this linear transformation is shown below(figure 5.)



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 rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel neighborhood for each
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

Fig. 5. Predicted Image

We have shown here what we able to clean through this model to show the power of simplest modelling technique available in machine learning. In further section we will consider complex scenarios where background is darker than writing and stained with coffee and ink spots. We will justify that indeed combining all the models having sparse features results in much more reduced error than using just one model.

3.2. K- means Clustering

Its obvious that using simple linear model doesn't give much improvement, that's because simple linear model just able to adjust for brightness and contrast it doesn't model the characteristics of stains. First observation to remove stain is that writing is darker than background around it. This observation is useful for crease like noisy image as simple linear doesn't perfectly remove creases as seen from figure 5 the remaining crease lines are those which are shadows in figure 4. and the remaining crease lines are lighter and narrower than writing. Thus we hypothesize that pixel contains information of surrounding pixels. So we can use clustering technique to cluster the pixels into writing and non-writing one's so that once we know the value of pixel below which all are non-writing one's then we can threshold the image. Thresholding is the process to cut out the portion so we are remains with the useful one. Manually we can do this by plotting the histogram of pixel brightness in a image shown in figure 6. and find the local maxima which creates a natural break in the pixel values. From the figure 6. these are 0.3 and 0.65. But instead of doing it manually it can simply be automated for each of the different images using clustering technique[45,41] to find out the lower and higher threshold.

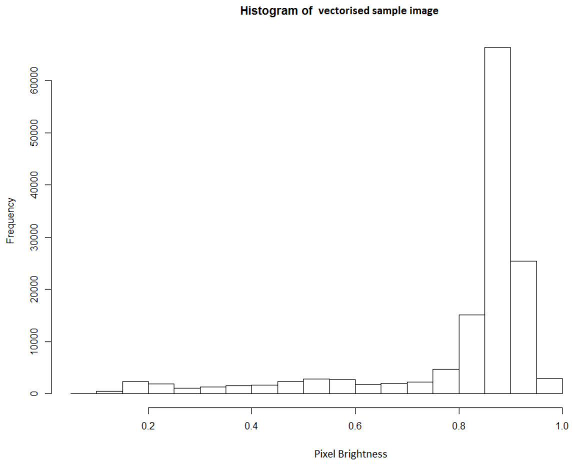


Fig. 6. Histogram for Image as vector

3.2.1. K-means Clustering Thresholding

Assuming that Z = M x N be the size of image, and let K be the number of clusters in which we want to divide the whole document, in our case k=3 as we want to cluster simple background, complex background and text regions. So for thresholding the non-text region the algorithm is explained as follows

- Let $[C_1 C_2 C_3 ....C_K]$ be the K points as initial centroids arbitrarily chosen from among the Z points.
- Now compute the distance of each point from the center of each class. Let the points be represented as X(i) where i varies from 1 to Z. So the euclidean distance can be calculated as

  $$d_k^{(i)} = \sqrt{\sum(X(i) - C_k{}^2)} \qquad (1)$$

  where k varies from 1 to K and i = 1 2 3 .... N.
- If $d_K = \min_{1 \leqslant k \leqslant K} [d_k^{(i)}]$ then X(i) belongs to cluster C_K.
- When all the points are assigned to K clusters recalculate the cluster centers as follows

  $$C_k^{new} = \sqrt{\frac{1}{N} \sum_{X^{(i)} \in C_k} X^{(i)}} \qquad (2)$$

- Now if $\sqrt{\sum(C_k^{new} - C_k)^2} \leqslant \Delta$, Δ is taken as 0.2 is satisfied then the iteration stops other wise take these new cluster centers as initial centers reiterate the steps 2 and 3 until the Δ is satisfied between initial and new cluster centers.
- The $j_t h$ threshold being calculated as $t_j = 0.5(C_j + C_{j+1})$ where $1 \leqslant j \leqslant K - 1$.

Using the algorithm as explained above again the threshold values come out to be same 0.3 and 0.65. Lets see the effect of thresholding using both thresholds each time in figure 7 and 8.

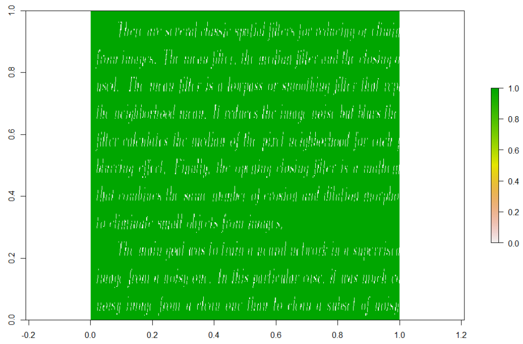

There are several classic spatial filters for reducing or elimin
from images. The mean filter, the median filter and the closing o
used. The mean filter is a lowpass or smoothing filter that rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel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

Fig. 7. Thresholding at 0.3

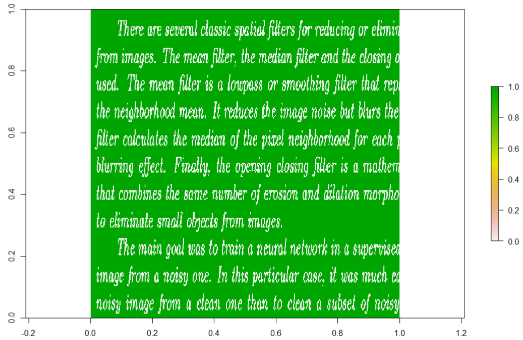
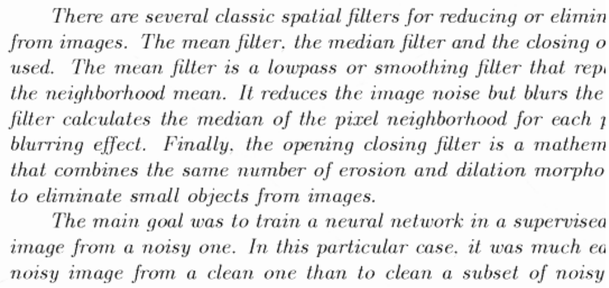

There are several classic spatial filters for reducing or elimin
from images. The mean filter, the median filter and the closing o
used. The mean filter is a lowpass or smoothing filter that rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel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

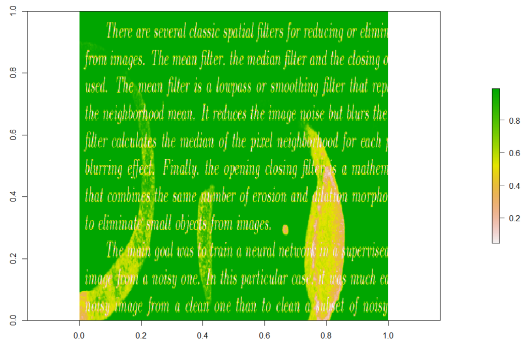
Fig. 8. Thresholding at 0.65

Thresholding at 0.3 removes out lot of writing but it removes the stains also while thresholding at 0.65 leaves writing but removes lots of stains as shown in figure 7 and 8. Obviously using 0.65 as threshold is more meaningful as it remove most of the stain while retain the text. Thus in this section we gained success in remove any crease mark from the paper. Combining the previous model and this model using gradient boosting machines results in RMSE improvement to 6.48%. Figure 9 shows the sample image after ensembling but it is intentionally shown what we have been able to clean at this point the figure 10 shows the sample image which we don't able to clean as it contains the darker spot which we didn't model upto this stage.



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 rep
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 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

Fig. 9. Cleaned image after ensemble



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 rep
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 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

Fig. 10. Image stained with coffee cant be cleaned

In further sections we consider more complex stains to model so that each of the model feature are sparse and non-overlapping thus giving us more improvement than possible using single model.

3.3. Image Morphology

As thresholding can't be able to separate the writing from the more complex stains like coffee spots, ink spots so have to engineer a feature that can characterizes these type stains. As these stains surround writing in wide region it can be hypothesize that we have to separate writing from the stained local maxima which can be easily done by a edge detector, canny edge detector is one such popular algorithm it runs as follows -

- **Smooothing** : As the image may contain some noise it applies gaussian filter with $\sigma = 1.4$ to remove noise which is like blurring the image.
- **Gradient Calculation** : The edges are determined where the gradient changes the most, this can be determined by finding the gradient of image at each pixel by applying the well known Sobel Operator. The kernel to find the gradient in x and y direction are -

  $$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \text{ and } K_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

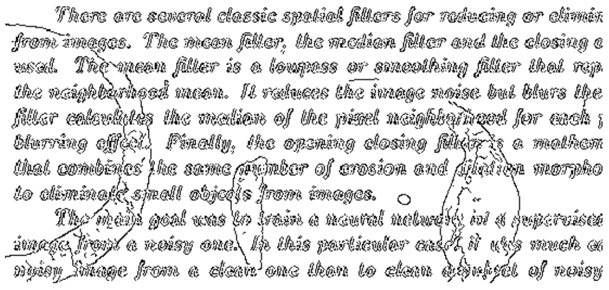
  Then the gradient magnitude is determined as the euclidean distance measure

  $$\Delta = \sqrt{\Delta_x^2 + \Delta_y^2}$$

  and direction as the edges are thick it requires to accurately place the edge

  $$\theta = tan^{-}1\left(\frac{|\Delta_y|}{|\Delta_x|}\right)$$

- The last step is to detect the strong edges and convert the blurred edges to sharp edges. Compare the pixel value in a 8 x 8 region to find the maximum pixel value replace with that value.



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 rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel neighborhood for each
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 supervise
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

Fig. 11. Canny Edge Detector

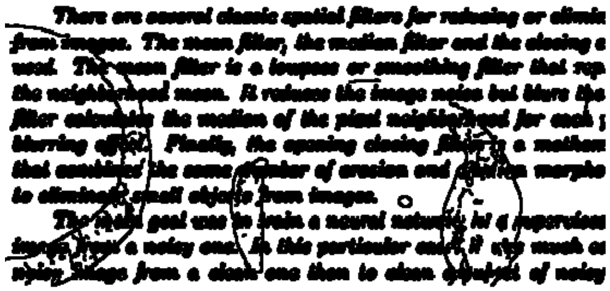
The sample image after edge detection is shown in figure 11. It has been observed from the image that there are edges which surround the writing and also the stains there are very less edges inside the stain. Furthermore the edges surround the writing occur in pairs while stain is surrounded by only boundary edge. Thus we can use this feature to remove stains from writing. Erosion and dilation are two such techniques of image morphology which are helpful in this case, as dilation in general is thought of making edge thicker by one pixel and erosion is just opposite removing a one pixel thick layer from the boundary of an object. On dilation the edges of writing expand to fill the empty space between them while the stain only grows by one thick layer as compared two writing which grows by two pixels. So if we do erosion the edges it will shrink the edges but finally we have some improvement in writing as compared to stain if we have successive dilation and erosion the result can be observed from the sample image in figure 12. Mathematically dilation can be defined on image mask I such that it connects the text edges to cluster them. The number of pixels it grows depends on the size of structuring element E. The dilation operation is given by equation 3.

$$D = I \oplus E = \left\{ h \in Z^2 | h = x + b, x \in I, b \in E \right\} \quad (3)$$

whereas the erosion can be defined as in equation 4.

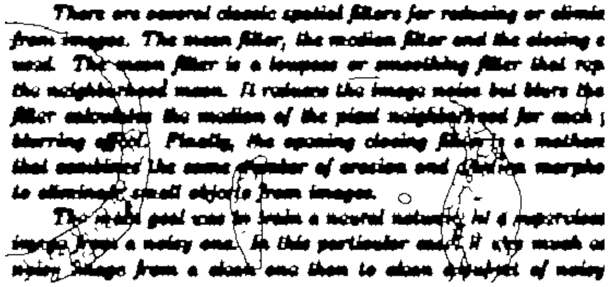
$$E = I \ominus E = \left\{ h \in Z^2 | x + b \in I \;\; for \;\; every \;\; b \in E \right\} \quad (4)$$

Here we have used a structuring element of size $[4x4]$. Figure 13. shows the sample image after once we applied erosion and dilation results in darkening the text and thining the stain,if we succesively do this for three time we find the stain is completely removed as shown in figure 14.



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 rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel neighborhood for each
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 supervise
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

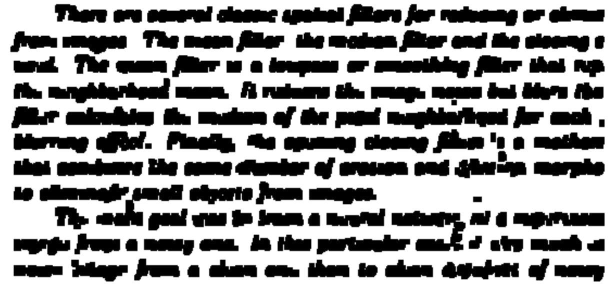
Fig. 12. Image Morphology : Dilation



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 rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel neighborhood for each
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 supervise
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

Fig. 13. Image Morphology : Erosion

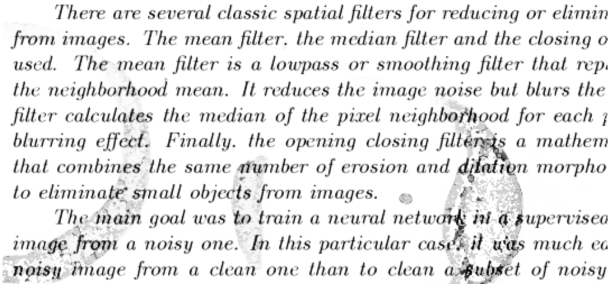
This works good as all of the writing is black, but only a small part of the stain remains. the stain has a thin line, while the writing has thick lines. So we can erode once, then dilate once, and the thin lines will disappear.The stain is now almost completely removed.



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 rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel neighborhood for each
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 supervise
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

Fig. 14. Successive dilation and Erosion(Three times)

Let's put it all together with the existing features that we have developed, by adding canny edges and the dilated / eroded edges to the gradient boosted model. This improved the RMSE score on the training dataset from 6.48% to 4.1%. The predicted sample image after ensembling is shown in figure 15. although we have improved much but we have to further process the image for cleaning it and reducing the RMSE.



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 rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel neighborhood for each
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 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

Fig. 15. Ensemble model with Image morphology

3.4. Background Removal

In the previous section even we have model the stain structure we can't be able to remove the stain completely. So we have to devise another feature which is non-overlapping with feature of previous section as well as more robust to any type of stain. From the previous observation it has been seen that stain manly comprises of the large part so it must surely be part of background. Considering the methods of extracting the background of an image we particularly known that taking moving average of a video stream gives out a background image. So we have to devise a filters which moves out across the image to extract the background, one such filter is the median filter. Median filter of width w is an image filter which moves out across the whole image while moving it replace the pixels surrounding it by the median of pixels on size of w x w, lets say the part of original dirty image for the size w x w is denoted by DI(w) whereas background image be denoted by BI(w) so the background removal be formulated as given in equation no. 5

$$BI(w) = median(OI(w)) \tag{5}$$

subject to $|BI(w) - OI(w)| \leqslant T$ where T is the threshold we taken as 0.1 and the width w is found out using the grid search method is taken to be 17 in this paper. The background image is shown in figure 16.

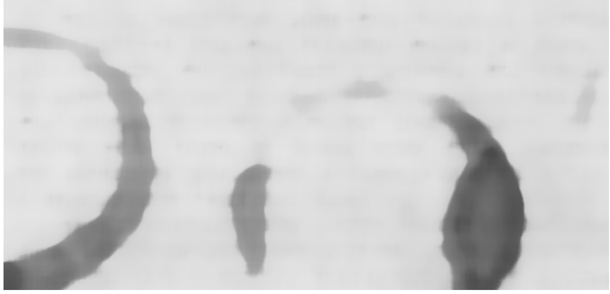
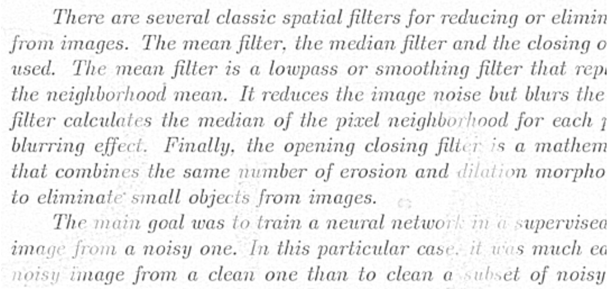


Fig. 16. Background Image

Clearly seen from figure 16. it contains the coffee cup stains and also the shade of the paper upon which the writing appears. So if we remove that part from the original image we are left with mostly writing although some writing may also be removed because it retains only those pixels which are darker than writing as shown in figure 17., that we don't have to care because finally we ensemble all the models that take care of this unwanted removal.

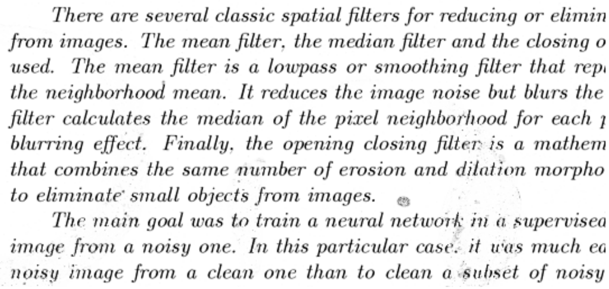


There are several classic spatial filters for reducing or elimin
from images. The mean filter, the median filter and the closing o
used. The mean filter is a lowpass or smoothing filter that rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel neighborhood for each
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 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

Fig. 17. Background Removal

We are using GBM package to create a predictive model. But as we have added more predictors or features, it has started to take a long time to fit the model and to calculate the predictions. So we have to switch to xgboost package.The final sample image after ensembling is shown in figure 18.



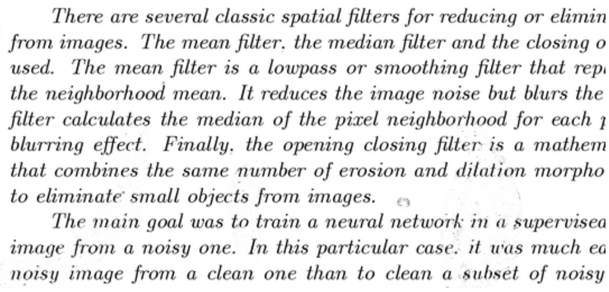
There are several classic spatial filters for reducing or elimin
from images. The mean filter, the median filter and the closing o
used. The mean filter is a lowpass or smoothing filter that rep
the neighborhood mean. It reduces the image noise but blurs the
filter calculates the median of the pixel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

Fig. 18. Ensemble Model to include Background Removal

The overall result represents an improvement in RMSE score on the training data from 4.1% to 2.4%.

3.5. *Sliding Window*

Considering the progress made so far it has been observed that the more we take into account the surrounding pixels the more robust the model is, so the most obvious modelling approach is to use a sliding window of size w x w after suitably padding the image with p pixels surrounding it and take these w^2 pixels as a predictor for the pixel in the center of the sliding window. In this paper we have used w=5 and p=3 and the predictor is random forest. To visualize the importance of these nearby pixels as a separate feature see the feature importance graph in figure 20. Sample image after ensembling with XGBoost is shown on the figure 19.



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 rep
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 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

Fig. 19. Ensemble Model to include Sliding Window Features

Clearly the sample image is much more cleaned as compare to previous modelling approaches.Overall improvement can be verified through the RMSE scores. RMSE on the training data dropped from 2.4% to 1.4% for this model.

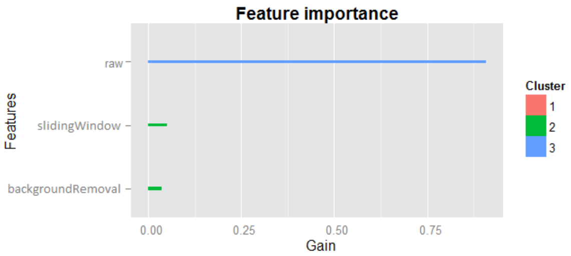


Fig. 20. Feature Importance

3.6. *Sparse Autoencoder*

Autoencoder consisits of encoder and decoder. Mainly the work of encoder is to transform the input image into some other hidden representation through the use of some non-linear function such that when this representation is used to transform to original image it does perfectly be able to reconstruct the image. This means that either the input image may contain some redundant or least significant components which do not provide any necessary information about the input representation or it contains noise which is being filtered to form a hidden representation which learns by itself as what parts are unnecessary we exactly require this thing only in our case. Let observe mathematically what it does. Let the input be $X \in R^p$ and let the encoder weight matrix be $W_e \in R^{sxp}$ and $d \in R^s$ be the bias of encoder then the transformation can be written as $H = g(W_eX + b)$ where g(.) is non-linear activation function for hidden representation. The decoder performs exactly opposite to encoder for the $H \in R^s$ derived from encoder it performs for decoder weight matrix $W_d \in R^{pxs}$ and bias $b \in R^s$ it gives back $\widehat{X} = g(W_dH + b)$. So the main task is to reduce the mean square error between $\widehat{X}$ and X. But for the perfect reconstruction and to learn the same feature it is required that $W_e = W_d^T$. the cost function is written as :

$$J = \min_{W,b,c} \left\| \widehat{X} - X \right\|_2^2 + \Psi(H) \qquad (6)$$

$$= \min_{W,b,c} \left\| f(W_d^T f(W_e^T X + b)) - X \right\|_2^2 + \Psi(H)$$

Further in the equation 6. there is regularization denoted by $\Psi(.)$ on the hidden representation which can be done by either of the techniques like drop-out, early-stopping to learn the sparse representation[50,54] the advantage of which is obvious in our case to reduce the no. of redundant computations.But this seems like overdoing the things as we have the training set available with cleaned images for comparison then it does make any sense to include the decoder for backpropagate the error to train the network so we will use the truncated version of Autoencoders that will be described in the next section.

3.7. Truncated Sparse Convolutional Autoencoder(TSCA)

As the Autoencoders are likely to handle the vectorised form of data and we have images which are stored in the form of matrices it is more conceptual to use convolutional AE because the convolutional neuaral nets[47,48,49] are motivated or developed for the vision problem. Since this is a vision problem of recognizing text from the noisy background it is more suitable to use convolution neural networks but as we describe in the previous section it does not need to reconstruct the image using convolutional decoder as we have have the clean image available so what we just want is the input the dirty image and output the cleaned one. Specifically, for an image $X \in R^{nxm}$, it encodes it by convolving it with encoder filters $W_e \in R^{sxsxd}$, resulting in d feature maps. These feature maps are then passing through a nonlinear activation function to generate their corresponding feature maps $Z \in R^{(n-s+1)x(m-s+1)}$, $i = 1....d$. Training the truncated AE can be implemented as given in equation 7 and 8.

$$J = \min_{W,b,c} \sum_{l=1}^{L} \left\| \widehat{X}^{(l)} - X^{(l)} \right\|_2^2 \qquad (7)$$

where

$$\widehat{X^{(l)}} = H_{p,s}(\widehat{Z}^{(l)}) = H_{p,s}(f(W_e * X^{(l)} + b_i)) \qquad (8)$$

where $H_{p,s}(Z)$ is the sparsifying operator it first max-pools and unpool with pooling size p and stride s. In the convolutional Autoencoders there are first few layers repeatedly apply the same weights across overlapping regions of the input data. Intuitively this is like applying an edge detection filter where the network finds the appropriate weights for several different edge filters.We have used 3 hidden convolutional layers, each with 25 image filters. The network architecture used in this paper is shown in figure 21. In this paper we have used SGD algorithm to train the network. Comparing the time complexity with the current work in this scenario DN[47] time complexity is $\mathscr{O}(d*s*(nm))$ while that of while the ConvSC[49] is given as $\mathscr{O}(d*(nm))$ whereas Truncated convolutional sparse AE does it in $\mathscr{O}(d*(kn))$ as k¡m it justifies the improvement in time complexity. During the first several iterations, the neural network is balancing out the weights so that the pixels are the correct magnitude, and after that the learning begins. Some of the convolutional filters are shown in figure 22. Figure 23. shows the loss function for both the training and test loss we have used SGD(Stochastic Gradient Descent) in this paper, it has been observed that the loss function started to decrease t=in the fist 50 iterations only and converges to the 0.001 error within 150 iterations this is the motivation why we have used the sparse structure not only to increase the accuracy but also reduce the training time and better generality and obviously less complex network.

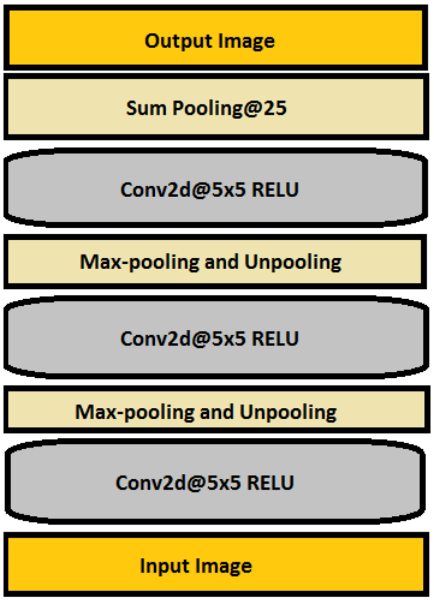


Fig. 21. Network Architecture

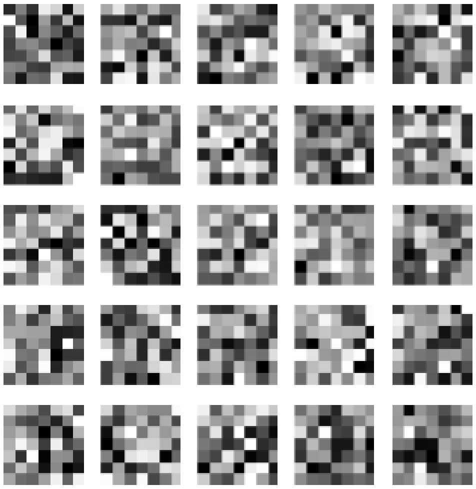


Fig. 22. Convolutional Filters

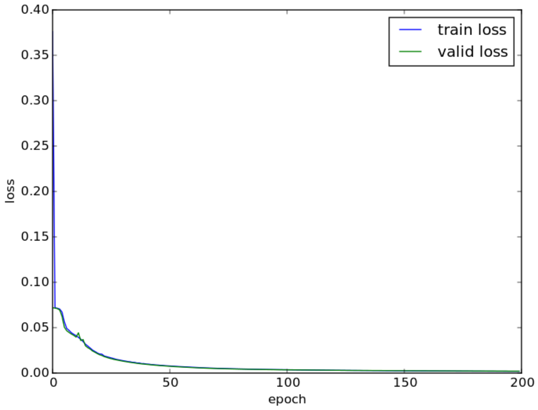
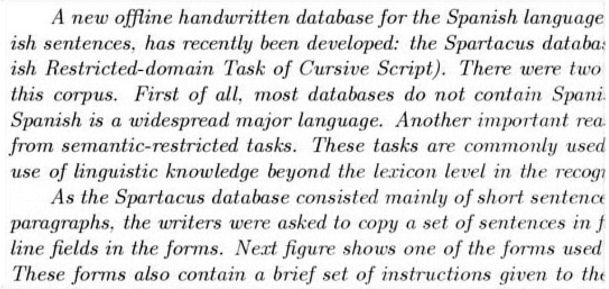


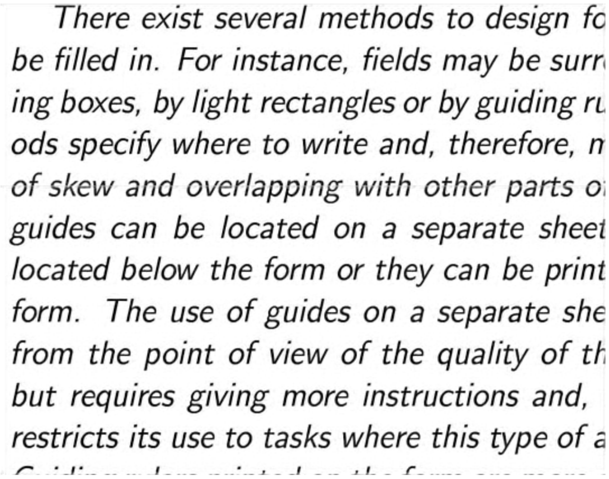
Fig. 23. Loss function

As seen from figure 22. filters are like some part of text letters that's actually we want the filters to do i.e. recognizing the text from the image to clean it. Figure 23. shows the loss function of neural network as seen loss function converges within 150 iterations of SGD although we have run upto 200 iterations. Figure 24 shows the cleaned image after TSCA to further improve on the quality and RMSE ensembled model predicted image is shown in Figure 25. shows the cleaned image after ensemble all the models clearly the predicted image is cleaned one without any stains thus we achieve for what we wanted. RMSE at this stage has dropped down to 0.7% which is clearly a significant improvement just half from the previous ensemble.



A new offline handwritten database for the Spanish language
ish sentences, has recently been developed: the Spartacus databa:
ish Restricted-domain Task of Cursive Script). There were two
this corpus. First of all, most databases do not contain Spani.
Spanish is a widespread major language. Another important rea
from semantic-restricted tasks. These tasks are commonly used
use of linguistic knowledge beyond the lexicon level in the recogn
As the Spartacus database consisted mainly of short sentence
paragraphs, the writers were asked to copy a set of sentences in f
line fields in the forms. Next figure shows one of the forms used
These forms also contain a brief set of instructions given to the

Fig. 24. Processed image from TSCA



There exist several methods to design fo
be filled in. For instance, fields may be surr
ing boxes, by light rectangles or by guiding ru
ods specify where to write and, therefore, m
of skew and overlapping with other parts o
guides can be located on a separate sheet
located below the form or they can be print
form. The use of guides on a separate she
from the point of view of the quality of th
but requires giving more instructions and,
restricts its use to tasks where this type of a

Fig. 25. Ensemble to include TSCA

4. Results

Apart form the improvement shown on the images of dataset, the theoretical measure to show the performance improvement we have used in this paper is RMSE(Reduced Mean Squared Error). Lets say the original clean image in the training dataset be denoted by X and the predicted image be denoted by $\widehat{X}$. Then the mathematical formula to calculate the RMSE is given in equation 9

$$RMSE = \sqrt{\mathbb{E}((\widehat{X} - X)^2)} \tag{9}$$

$$= \sqrt{\sum_{i,j} (\widehat{x}_{ij} - x_{ij})^2 / N}$$

where $N =$ width x height of image. The RMSE at each step of whole model is summarized in the following table.

RMSE	
Models	RMSE
Linear Model	7.78%
Linear Model + Clustering	6.48%
All previous + Image Morphology	4.1%
All previous + Background Removal	2.4%
All previous + Sliding Window	1.4%
All previous + TSCA	0.7%

5. Conclusions

This paper proposed a novel and effective method for cleaning stained document. The method has been proved to be efficient for reducing the RMSE to less than 1%. The proposed method applies many techniques to model the text and non-text regions in the image it has the advantage of being generalize method over other methods described in the literature review it require training of the model to any arbitrary complex noisy images and one such publicly available dataset Is NoisyOffice provided by UCI machine learning repository[14]. The method here explore the previous approaches like K-means clustering thresholding, Image Morphology, Sliding Window and Background Removal but finds that any method by itself is not robust neither complete to remove the stains from the documents. Further we take care of the case that each of the models described have features which are sparse or non-overlapping enough to show the remarkable improvement in the RMSE scores. This has been showed at each step to demonstrate the improvement and cleaning of the documents at each step. Otsu's method which is manual way of thresholding the image we have automated it using the K-means clustering. As the background removal removes the text region we take care of that by ensembling it with other models and taking also the nearby pixels as a feature marks the significant improvement over most of the methods. Training the CNN takes long time but incorporating the sparse structure into it we just achieved what we wanted within 150 iterations with significant reduction in error, Although the cleaned image after TSCA is very cleaned but it is only after ensembling the image becomes so easy to see we can't observe the difference between cleaned image and predicted one this is because the very first model in our ensemble that is linear regression takes care of contrast and brightness of the image. Finally we see towards experimenting and deploying this system for the real world applications.

Acknowledgments

It is to thanks all the professors of Electrical Engineering Department Of Indian Institute of Technology, Delhi for providing the enough support towards learning the concepts and encourage to do real world application based projects out of this learning. Also I would like to thanks all my friends to encourage me to write about this project so that it can be used further to implement in real time system.

Conflicts of Interest: The author declares that there is no conflict of interest regarding the publication of this paper.

References

1. U. Schmidt, Q. Gao, and S. Roth, A generative perspective on MRFs in low-level vision, in CVPR, 2010.
2. B. A. Olshausen and D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, vol. 381, pp. 607-609, 1996.
3. H. Lee, A. Battle, R. Raina, and A. Y. Ng, Efficient sparse coding algorithms, in NIPS, pp. 801-808, 2007
4. B. A. Oishausen, C. H. Anderson, and D. C. V. Essenla, A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information, J Neurosci, vol. 13, pp. 4700-4719, 1993.
5. L. Wiskott, J.-M. Fellous, N. Krger, and C. von der Malsburg, Face recognition by elastic bunch graph matching, PAMI, vol. 19, pp. 775-779, 1997.
6. B. J. Frey and N. Jojic, Transformation-invariant clustering using the EM algorithm, PAMI, vol. 25, pp. 1-17, 2003.
7. D. B. Grimes and R. P. N. Rao, Bilinear sparse coding for invariant vision, Neural Comp, vol. 17, pp. 47-73, January 2005.

8. C. K. I. Williams and M. K. Titsias, Greedy learning of multiple objects in images using robust statistics and factorial learning, Neural Comp, vol. 16, pp. 1039-1062, 2004.
9. B. J. Frey and N. Jojic, A comparison of algorithms for inference and learning in probabilistic graphical models., PAMI, vol. 27, pp. 1392-1416, Sept. 2005.
10. J. Winn and A. Blake, Generative affine localisation and tracking, NIPS, 2004.
11. J. Lucke and J. Eggert, Expectation truncation and the benefits of preselection in training generative models, JMLR, vol. 11, pp. 2855 - 2900, 2010.
12. M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, An introduction to variational methods for graphical models, Mach Learn, vol. 37, pp. 183-233, 1999.
13. L. Shen and L. Bai, A review on Gabor wavelets for face recognition, PAA, vol. 9, no. 2-3, pp. 273-292, 2006.
14. Lichman, M. (2013), UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
15. Y. LeCun, K. Kavukcuoglu, and C. Farabet, Convolutional networks and applications in vision, in ISCAS, pp. 253 - 256, 2010.
16. N. Jojic, B. J. Frey, and A. Kannan, Epitomic analysis of appearance and shape, in ICCV, pp. 34-41, 2003.
17. R. G. Casey and E. Lecolinet, A survey of methods and strategies in character segmentation, PAMI, vol. 18, pp. 690-706, 1996.
18. P. Dayan and R. S. Zemel, Competition and multiple cause models, Neural Comp, vol. 7, pp. 565 - 579, 1995.
19. D. G. Lowe, Distinctive image features from scale-invariant keypoints, vol. 7, IJCV, vol. 60, pp. 91-110, 2004.
20. N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in CVPR, 2005.
21. H. Zhang, K. Zhao, Y.Z. Song and J. Guo, Text extraction from natural scene image: A survey, Neurocomputing 122 (2013) 310-323.
22. K. Jung, K.I. Kim and A.K. Jain, Text information extraction in images and video: a survey, Pattern Recognition 37 (2004) 977-997.
23. X. Liu and J. Samarabandu, An Edge-based Text Region Extraction Algorithm for Indoor Mobile Robot Navigation, in: Proceedings of the IEEE International Conference on Mechatronics & Automation, IEEE (Niagara Falls, Canada, 2005), pp. 701-706.
24. C. Liu, C. Wang and R. Dai, Text Detection in Images Based on Unsupervised Classification of Edge-based Features, in: Proceedings of the 8th International Conference on Document Analysis and Recognition, IEEE Computer Society (2005), pp. 610-614.
25. M.R. Lyu, J. Song and M. Cai, A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction, IEEE Transactions on Circuits and Systems for Video Technology 15 (2005) 243-255.
26. T.N. Dinh, J. Park and G. Lee, Low-Complexity Text Extraction in Korean Signboards for Mobile Applications, in: 8th IEEE International conference on Computer and Information Technology, IEEE (Sydney, NSW, 2008), pp. 333-337.
27. A.N. Lai and G. Lee, Binarization by Local K-means Clustering for Korean Text Extraction, in: IEEE International Symposium on Signal Processing and Information Technology, IEEE (Sarajevo, 2008), pp. 117-122.
28. S. Grover, K. Arora and S.K. Mitra, Text Extraction from Document Images using Edge Information, in: Annual IEEE India Conference, IEEE (Gujarat, 2009), pp. 1-4.
29. T.Q. Phan, P. Shivakumara and C.L. Tan, A Laplacian Method for Video Text Detection, in: 10th International Conference on Document Analysis and Recognition, IEEE Computer Society (Barcelona, 2009), pp. 66-70.
30. P. Shivakumara, T.Q. Phan and C.L. Tan, Video text detection based on filters and edge features, in: IEEE International Conference on Multimedia and Expo, IEEE (New York, 2009), pp. 514-517.
31. X. Zhang, F. Sun and L. Gu, A Combined Algorithm for Video Text Extraction, in: 7th International Conference on Fuzzy Systems and Knowledge Discovery, IEEE (Yantai, Shandong, 2010), pp. 2294-2298.
32. H. Anoual, D. Aboutajdine, S.E. Ensias and A.J. Enset, Features Extraction for Text Detection and Localization, in: 5th International Symposium on I/V on Communications and Mobile Network, IEEE (Rabat, 2010), pp. 1-4.
33. S. Shah, C. Modi and M. Patel, Novel Approach for Text Extraction from Natural Images Using ISEF Edge Detection, in: International Conference on Emergingtrends in Networks and Computer Communications, IEEE (Udaipur, 2011), pp. 487-491.
34. L. Zheng, X. He, B. Samali and L.T. Yang, An algorithm for accuracy enhancement of license plate recognition, Journal of Computer and System Sciences 79 (2013) 245-255.
35. J.L. Yao, Y.Q. Wang, L.B. Weng and Y.P. Yang, Locating Text based on Connected Component And SVM, in: Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition, IEEE (Beijing, China, 2007), pp. 1418-1423.
36. Y.F. Pan, X. Hou and C.L. Liu, Text Localization in Natural Scene Images based on Conditional Random Field, in: 10th International Conference on Document Analysis and Recognition, IEEE Computer So-

ciety (Barcelona, 2009), pp. 6-10.

37. W. Kim and C. Kim, A New Approach for Overlay Text Detection and Extraction From Complex Video Scene, IEEE Transactions on Image Processing 18 (2009) 401-411.
38. L. Sun, G. Liu, X. Qian and D. Guo, A Novel Text Detection and Localization Method based on Corner Response, in: IEEE International Conference on Multimedia and Expo, IEEE (New York, 2009), pp. 390-393.
39. M. Kumar, Y.C. Kim and G.S. Lee, Text Detection using Multilayer Separation in Real Scene Images, in: 10th IEEE International Conference on Computer and Information Technology, IEEE Computer Society (Bradford, 2010), pp. 1413-1417.
40. Y. Zhang, C. Wang, B. Xiao and C. Shi, A New Text Extraction Method Incorporating Local Information, in: International Conference on Frontiers in Handwriting Recognition, IEEE (Bari, 2012), pp. 252-255.
41. Ursula Gonzales-Barron, Francis Butler, A comparison of seven thresholding techniques with the k-means clustering algorithm for measurement of bread-crumb features by digital image analysis, In Journal of Food Engineering, Volume 74, Issue 2, 2006, Pages 268-278, ISSN 0260-8774.
42. X.W. Zhang, X.B. Zheng and Z.J. Weng, Text Extraction Algorithm under Background Image using Wavelet Transforms, in: Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition, IEEE (Hong-Kong, 2008), pp. 200-204.
43. Z. Ji, J. Wang and Y.T. Su, Text Detection in Video frames using Hybrid features, in: Proceedings of the 8th International Conference on Machine Learning and Cybernetics, IEEE (Baoding, 2009), pp. 318-322.
44. T. Zhao, G. Sun, C. Zhang and D. Chen, Study on Video Text Processing, in: IEEE International Symposium on Industrial Electronics, IEEE (Cambridge, 2008), pp. 1215-1218.
45. D. Liu and J. Yu, Otsu Method and K-means, 2009 Ninth International Conference on Hybrid Intelligent Systems, Shenyang, 2009, pp. 344-349.
46. Z. Zhang, W. Shen, C. Yao and X. Bai, Symmetry-Based Text Line Detection in Natural Scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (Boston, 2015) pp. 2558-2567.
47. M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, Deconvolutional networks, in Proc. CVPR, 2010, pp. 2528-2535.
48. M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in Proc. ICCV, 2011, pp. 2018-2025.
49. K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mahieu, and Y. LeCun, Learning convolutional feature hierarchies for visual recognition, in Proc. NIPS, 2010, pp. 1090-1098.
50. M. Ranzato, Y.-L. Boureau, and Y. LeCun, Sparse feature learning for deep belief networks, in Proc. NIPS, 2007, pp. 1-8.
51. Y.C. Wei and C.H. Lin, A robust video text detection approach using SVM, Expert Systems with Applications 39 (2012) 10832-10840.
52. H. Xu and F. Su, A Robust Hierarchical Detection Method for Scene Text based on Convolutional Neural Networks, in: IEEE International Conference on Multimedia and Expo, IEEE (Turin, 2015), pp. 1-6.
53. K. Chen, F. Yin, A. Hussain and C.L. Liu, Efficient Text Localization in Born-Digital Images by Local Contrast-Based Segmentation, in: 13th International Conference on Document Analysis and Recognition, IEEE (Tunis, 2015) pp. 291-295.
54. J. Li, T. Zhang, W. Luo, J. Yang, X.-T. Yuan, and J. Zhang, Sparseness analysis in the pretraining of deep neural networks, IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 6, pp. 1425-1438, Jun. 2017