A New Approach to Robust, Quality-Independent Scene Text Processing: Using Bi-level overlapped binning and Delta Images

# ABSTRACT

Extracting textual information from scene text images has been an active field of research with lots of work done in areas of localization, recognition and binarization. These papers provide new features or new approaches to preserve text and remove non-text as much as possible and new features, techniques or machine learning models come out almost every few months that can beat the existing state-of-the-art. Though most of these papers have novel approaches, the first processing step on the raw image has remained nearly unchanged for a past decade. MSERs or one of it’s modified forms remain widely as the most used processing step for extracting regions from a scene text image. The new techniques are then applied on these extracted regions. Though MSERs are very useful, it often suffers from problems of accurate parameters and thresholding. Also, MSERs focus on extracting a single stable region in the image space which often leads to problems in low quality images where a stable region might not correspond to a text region but overlap with both text and a large portion of the background. In this paper, we present a novel new approach for scene image processing called bi-level overlapped binning which instead of giving a single solution for a homogenous region in a scene, outputs several solutions in the form of different connected components of binary images which we call Delta Images. These connected components represent regions in the scene text with different deviations of RGB values. It’s a divide-and-conquer based solution and we process each of the connected components individually. Finally, we recombine these solutions to achieve our desired results. In this paper we show the working of this new approach in an active field of research – Text Binarization and give our results on several standard datasets. The technique is able to extract text regions even from low quality images and has shown to be a very robust processing step for raw scene images

# The Overview

1. **Foreground – Background Pixels split**: The first step involves identifying and splitting the pixels of the image into sets of candidate foreground and candidate background pixels. We use a bi-level overlapped binning approach which gives different sets of candidate foreground and background splits. This is the critical step as the next set of steps depends on the fact that there is at least one split where the text pixels fall into one set of pixels and with very little background pixels in the same set.
2. **Homogeneous region detection:** Step 1 gives several sets of candidate foreground and background pixels. This step involves identifying clusters of connected pixels which represent a homogenous region in the given input image. A homogeneous region is defined as a set of connected pixels in an image whose deviation in RGB values is less than the deviation from the surrounding pixels of that cluster.
3. **Text/Non-Text Separation:** The homogeneous regions extracted in the form of connected set of pixels from Step 2 are now classified into text and non-text using a supervised learning model. Regions classified as non-text are removed and candidate text regions are retained for combination
4. **Probabilistic Combination:** Step 3 gives us multiple sets of candidate text pixels across multiple bin sizes used in the bi-level binning step. We compute the probability of a pixel being a text pixel based on its occurrence across the different delta images. We label pixels in the final binarized image based on a probability outcome
5. **Foreground – Background Pixels split: Bi-level Binning**

The first target is to identify sets of foreground and background pixels and split them. We use bi-level overlapped binning and adjacent bin combination which gives us a new set of binary images as output called Delta Images which have a unique property of segmenting the raw image into regions based on deviation of the region. The main aim of binning is to be able to capture all regions in an image whose deviation is less than a particular value, that value for a bin being called the bin size, in the form of single connected component such that geometric properties and other features can be calculated and text/non-text separation performed.

1. **Bi-Level Overlapped Binning**

For a given Bin Size S, we divide the entire range of RGB Euclidian distances from origin into two levels of Bins

*For a given Bin size S:*

**Level 1** (Bin number 1 to m)

*{(0 to s-1), (s to 2\*s – 1), (2\*s to 3\*s – 1) …., (m-1\*s to m\*s)}*

*OR*

**Level 2** (Bin number m+1 to 2\*m -1)

*{(0.5\*s to 1.5\*s – 1), (1.5\*s to 2.5\*s –1), …..., ((m-1.5) \*s to (m-0.5) \*s)}*

*OR*

*where an element (X to Y) denotes a bin such that only if a point has values between X and Y both inclusive, it is considered a data point for that bin and the pixel is labelled as 1 otherwise it is labelled as 0.*

*Where* is the minimum number of bins needed of size S such that the entire range of values possible for point falls into at least one bin’s range of values. *, is the maximum value possible for a pixel. Level 1 ranges from 0 to and Level 2 ranges from )*

Each Bin is represented as a binary image matrix which is the kth bin of size S and ()ij is the value of the ith row and jth column of

Where

R(i,j) = Red Value of pixel at ith row and jth column of image

G(i,j) = Green Value of pixel at ith row and jth column of image

B(i,j) = Blue Value of pixel at ith row and jth column of image

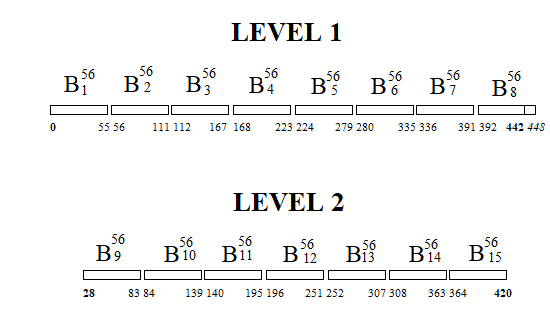


Fig 1. Bins for size = 56. The range corresponding to each Bin denotes the pixel of the corresponding binary image for the bin will be 1. If the value falls in that range then the binary image will have 1 for that pixel, otherwise it is 0.

For a given bin size s, there are total bins in both levels together where ,each represented by a binary image where the first images are the Level 1 bins and the next images are Level 2 bins. If we look at Fig 1 for a given bin at level 1, the bin at Level 2 overlapping with this bin on the left half is and with the right half is .

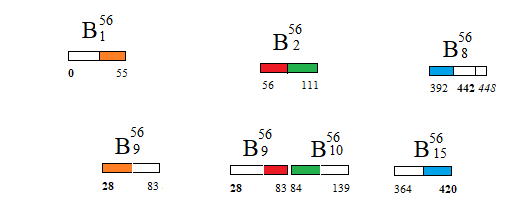


Fig 2. Diagram depicts the bins that overlap with the level 1 bins to the left and right. For bin no. 2 the right overlapping bin is and the left overlapping bin is . We see Bin no 9 overlapping to the left half and bin no. 10 overlapping to the right half of bin no. 2

In order to achieve robustness and quality independence, we use *Decremented Interval Binning*– using multiple bin sizes with the difference between corresponding bin sizes decreasing the higher we go. At this stage, let it be clear that when we say foreground or background pixel, it is for convenience and we do not label any specific set of pixels as foreground or background. Any one of the sets can correspond to either foreground or a background the next steps will do the work of eliminating the sets of background pixels.



Fig 3. A graph of different binary images to show how connected components behave on increasing bin size. The edges connect the subset bins to their superset bins(in subset/superset are in terms of PixelValue ranges each holds) at the immediate next bin size. The bin size denotes the size of the original Bins and the Range is the new size after the adjacent bin combination step

The lines in Fig 3 connect subset bins to their superset bins. For example, bin (374 to 406) is a subset of bin (367 to 420) in terms of range of pixel values and is connected by a line. Later, this relationship will be used to determine stable regions and extract features for text non-text separation.

1. **Adjacent bin combination**

Consider a Homogenous region Rwith *PixelValues* in the range of 30 to 60. We want all regions whose deviation is less than S to fall into at least one of the bins of size S such that the R occurs as a/part of a single connected component in a binary image and can be used for proper evaluation of structure and geometry for text/non-text separation. If suppose bin size is 32, then we want the *PixelValue*sto fall entirely in one of the bin thereby occurring as a/or part of a single connected component in a binary image. However, Let’s look at the *PixelValue* ranges of the first few bins of both the levels

Level 1

{(0 to 31), (32 to 63), (64 to 95), (96 to 127) ….}

Level 2

{(16 to 47), (48 to 79), (80 to 111) …}

We see in both the levels, the region R fails to fall into any of the bins despite having a pixel value deviation less than the particular bin size. The Bins underlined are the bins across which the region R is spread. It will basically occur divided into two parts with each part occurring at a different bin in both level 1 and level 2.

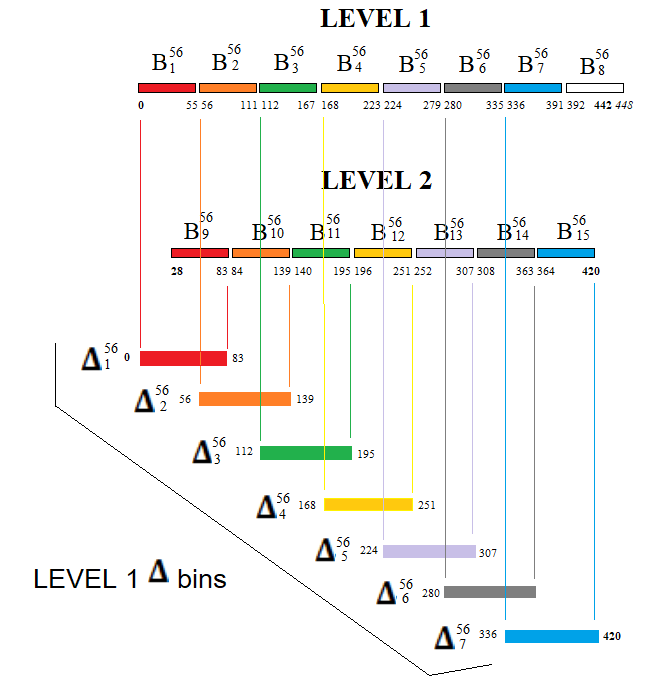
To solve this problem, we create a new set of recombined bins called as Δ bins, which are represented as Delta Images.

**LEVEL 1** Δ

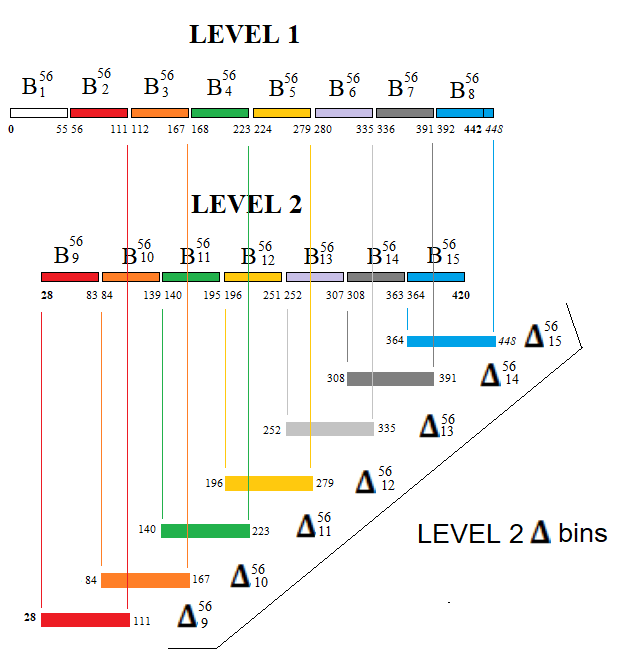
**LEVEL 2**

Where, +1

*Note: + operator while adding bins is pixel-wise logical OR operator of binary images. The results in another binary image called Delta Image.*



*Fig 4. An Illustration of the Level 1 Delta bins. The bins overlap in such a way such that all regions with deviation less than equal to 56 will always fall into one of the colour ranges(either at level 1 or at level 2 illustrated next) entirely (thus falling into one of the bins). Note*  does not have a Delta bin in level 1 but has a Delta bin in level 2 of the next figure.



*Fig 5. Illustration of Level 2 Delta bins for size 56. The range (colour block) corresponds to each Delta bin and denotes for what PixelValue of point (i,j) the binary image will be 1 at point (i,j). All other points of the binary image for a Delta bin will be 0 if it does not fall in the corresponding range*

Considering the previous example, Let us see the first few bins for bin size 32

LEVEL 1

{(0 to 47), (32 to 79), (64 to 111) ...…}

LEVEL 2

{(16 to 63), (48 to 95), (80 to 127) ….}

*Region R(range 30 to 60) falls entirely in 1st bin of Level 2 (range 16 to 63)*

The Delta bins have an important property which makes it useful for our task. Given a Delta Bin of combined from bins of size S (from henceforth will be referred as size of delta bin) a region in the image whose deviation is less than equal to S will occur as a/part of a single connected component in the same spatial coordinates as the region

We state the following lemma and its proof for any region R. *A region is a set of points on an image that are connected*

Lemma : *If the difference between the pixel with the highest PixelValue and lowest PixelValue in Region Ris D and let S be any bin size such that then if , ()ij = 1 for one value of positive integer*  +1

Proof: Let the minimum pixel value be for pixel (xa,ya) and maximum be for pixel (xb,yb) such that .

Let +1 ….. ***Def 1*** The range of *PixelValues* of the kth bin at level 1 contains

from the definition of ()ij ,

Now

Since,

***…… InEqn 1***

Also,

***..….. InEqn 2***

Thus, from *InEqn 1 and InEqn 2* ***………InEqn 3***

Since is an integer.

Implies that the pixels with value will fall in the same bin or the immediate next bin. Since Deviation is less than size of each bin the bins where the region R is spread over is limited to a single bin or two adjacent bins

Case 1: ***…… Eqn 4***

In that case, both from *Def 1* and from *Eqn 4*.

For any ,

Since is an integer and L.H.S is an integer. Only possible solution

ij

Now,

Since, ij for all

ij  = 1 for all

Case 2 : ***…… Eqn 5***

We have to consider two different sub cases depending on whether lies to the left half of the bin thereby overlapping with the left overlapping bin of level 2 or the right half of the bin thereby overlapping with the right overlapping bin of level 2

Case 2.1 : ***…… InEqn 4***

Let us consider bin no. , which is the bin that overlaps with the right half of bin k. +1 from lemma statement

Now,

From condition of *Eqn 6*,

**…… InEqn 5**

And from *Eqn 5* :

**…… InEqn 6**

***……. InEqn 7***

From Definition of ()ij

()XbYb = 1

Bin k has values ranging from and From *InEqn 4* Bin has the values

*Inference 1: Bin k and Bin k+m combined has all the pixel values P such that*

The pixel ranges Bin k and Bin m captures are overlapping. Since Pmin is in bin k and Pmax in bin m ,for any the following holds

**….InEqn 8**

From definition of

From *Inference 1* and *InEqn 5* we thus prove

ij  = 1 for all

case 2.2 :

, Since

Thus, ***……. InEqn 9***

+1

***……InEqn 10***

From *Def 1*,

***…….InEqn 11***

From *InEqn 6* and *InEqn 7*,

**…..InEqn 12**

From definition of ()ij

()XbYb = 1

Bin k+1 has values ranging from and from *InEqn 9 and InEqn 10* Bin k+m has the values

*Inference 2: Bin k+1 and Bin m combined has all the pixel values P such that*

The PixelValue ranges of Bin k+1 and Bin k+m are overlapping. Since Pmin is in bin m and Pmax in bin k+1,for any the following holds

**..InEqn 12**

From definition of

From Inference 2 and the *InEqn 9* we thus prove

ij  = 1 for all

*Thus for all the cases the lemma gets proved.*

During the proof we draw two important inferences, each of which gives us the range for a bin. The range of *PixelValues* is .

The Lemma tells us that if there is a text region and if the range of PixelValues in the text is D, all the pixels will occur as a/part of a connected component in a bin of size D (or more). Moreover, for the bin where it will occur, if the bin size is close to D as possible (best if it is D) then the chances of background pixels occurring in the same bin is also very less assuming that the deviation in PixelValues within a Text (or a homogenous region) is much less than the deviation between the Text pixels and the background pixels immediately surrounding the Text region. We find this to be true in nearly all the scene text images we come across.

Thus, the bins represented as binary images capture several homogenous regions for several different bin sizes. A connected component in one binary image of bin size S will have the same or more pixels for the next higher bin size.

1. **Homogenous Region Detection: Identifying stable regions using classifiers on connected components**

Step 1 gives several sets of connected components. All of the components do not correspond to homogenous regions in the original image. Bin sizes smaller than the deviation in a homogenous region will fragment the connected components into pieces. Fig 1 shows how a homogenous region fragments and gets distributed across several bins when bin size reduces and how the size of the connected component increases with increasing bin size. For a particular connected component, after a particular bin size, the size will stop increasing too rapidly on increasing the bin size to the next value and rate of increase of size of connected component will be minimum.

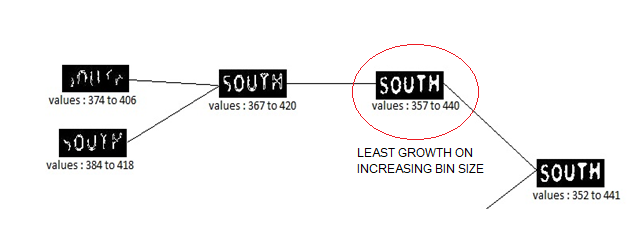
This property is used to extract un-fragmented homogenous regions. We compare rate of increase of the size of connected components when the range of PixelValues in a binary image increases. 

Fig 2: A part of the graph from figure 1. Along this path we see the components in the binary image representing bin (357 to 440) having the least growth on increasing bin size. Note, analysis is on each connected component of an image. Here incidentally one binary image has nearly all its components which have least growth on increasing bin size to the next value. This is a frequently occurring case since most real world text have nearly the same deviation for each letter

We evaluate stability of each connected components in a bin against two increases of the *PixelValue* ranges. The images compared against are called *Lower Range Incremented (*LRI) *Image* and *Higher Range Incremented* (HRI) *Image* where. The LRI of a Delta Image is basically a new image after adding the left overlapping bin of level 2 for level 1 Delta images and adding the left overlapping bin of level 1 for level 2 Delta images.

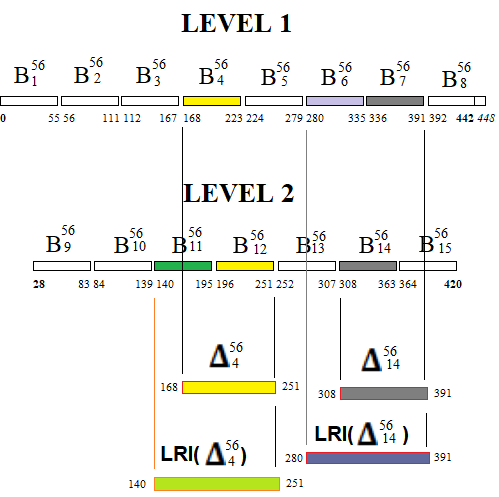


Fig 3. An Illustration of Delta Bins and their corresponding LRI Image ranges calculated by adding the left overlapping bin to the Delta bin.

Where +1

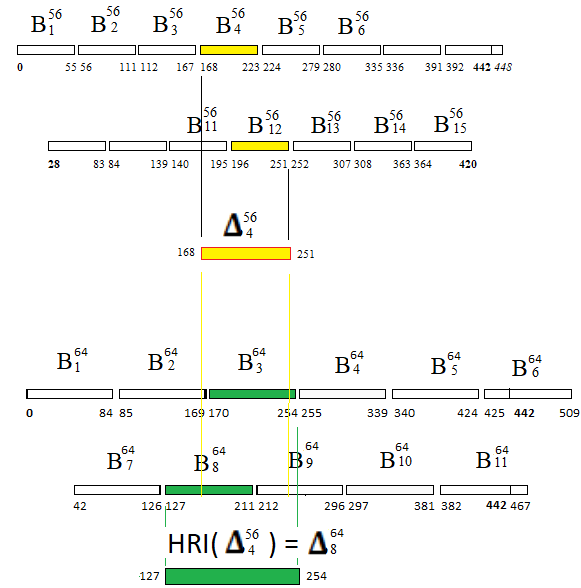


Fig 4. An illustration of HRI image range of *PixelValues* for given Δ Bin of size 56. The HRI image is a Delta Image corresponding to a Delta bin of size that is enough to cover the range of values of the Δ bin. HRI image is found by finding the Delta bin in the next higher bin size that is greatar than equal to such that the range of *PixelValues* of the lower bin sized delta bin is a subset of the range of *PixelValues* of the higher bin sized delta bin. The HRI image is the binary image corresponding to this bin. For example , has range(168 to 251) and occurs within the range of(158 to 251). The next bin size after 56 being 64 here.

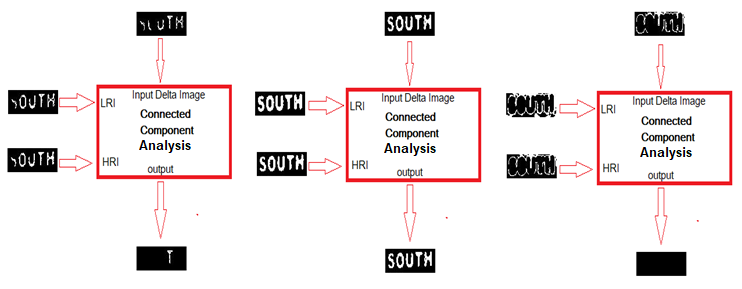


Fig 4. During connected component analysis we extract features from each connected component and decide whether it’s a stable text component or not. It preserves the connected component if it is a text region otherwise changes all the pixels to background for that component, effectively deleting it from any further processing. During Analysis we use classification which takes both stability features of section 2 and text features of section 3

For each connected component of the Delta Image (the binary image corresponding to a bin) we extract the following features and use it to classify stable and unstable regions. For each connected component C in the Delta image, we denote LRI(C) as the largest connected component of the LRI image that overlaps with C assuming the same 2D spatial coordinates for both images and HRI(C) as the largest connected component that overlaps with C in the HRI image

Features Set for Stable region detection used for classification

|  |  |
| --- | --- |
| Features | DESCRIPtiON |
|  | *No. of Pixels of connected component C* |
|  | *fractional increase in number of pixels of LRI(C)* |  |
|  | *fractional increase in number of pixels of HRI(C)* |
|  | *No. of Holes or gaps in C* |
|  | *fractional increase in number of holes of LRI(C)* |
|  | *fractional increase in number of holes of HRI(C)* |
|  | *Density of C* |
|  | *fractional increase in density of LRI(C)* |
|  | *fractional increase in density of LRI(C)* |
| 1. Bin Size S | *Bin size of the bin to which C belongs* |

From our training samples we find stable connected components will have small values for features 2, 3, 5, 6, 8, 9. Features 1, 4, 7, 10 have high to mid-range values in their respective domain space.

1. **Text/Non-Text Separation**

Features described in section 2 when used to train a classifier for separating text components from non-text ones removes nearly most of the unwanted connected components and connected components which do not cover a complete homogenous region in the image. In this step, we remove those homogenous components which are non-text (example the background, lines across image, fragmented connected components which appear stable across LRI and HRI, etc.). We use the following features from [?]

The Features for Text/Non-Text separation used in the classifier

1. **SW(Stroke Width):** Stroke width is a widely used characterness cue. We use SW of the connected components which is measured as

E*(l)* and Var*(l)* are SW mean and variance

respectively

1. **Edge Histogram of Gradients(eHOG):** eHOG measures gradient orientation at the edges of the histogram. It exploits a feature that text edge pixels occur in pairs with opposite gradient directions. Since we have connected components in a binary image, we do not need any edge detector method. The edge pixels are the pixels with values 1 which are adjacent to 0’s. For each connected component, the largest continuous chain of connected edge pixels is taken as the skeleton. The pixels on the skeleton are then divided into 4 types depending on the direction(or angle) between the pixel and the next pixel

*Type 1*: 0 *< θ* ≤ *π/*4 or 7*π/*4 *< θ* ≤ 2*π*, *Type 2*: *π/*4 *< θ* ≤ 3*π/*4,

*Type 3*: 3*π/*4 *< θ* ≤ 5*π/*4, and *Type 4*: 5*π/*4 *< θ* ≤ 7*π/*4.

For Text, the number of edge pixels in Type 1 will be close to that in Type 3, and so for Type 2 and Type 4. eHOG(r) is then measured as

where *wi (r )* denotes the number of edge pixels in Type *i* within region *r*

We also observed that if a connected component C is a text component, HRI(C) and LRI(C) are also text components with similar SW and eHOG values. We therefore have the following final set of features for text/non-text separation

Features Set for Text/Non-Text separation used for classification

|  |  |
| --- | --- |
| Feature | DESCRIPTION |
| 1. *SWT(C)* | *Stroke width of C* |
| 1. *EHOG(C)* | *eHOG of component C* |
|  | *Fractional decrease of SW of HRI(C)* |
|  | *Fractional decrease of SW of LRI(C)* |
|  | *Fractional decrease of eHOG of HRI(C)* |
|  | *Fractional decrease of eHOG of LRI(C)* |

We extract all the features mentioned in Section 2 and Section 3 from a connected component and use it to train a classifier which gives us 1 for text and 0 for non-text and other fragmented text components. The connected components that are classified as 0 are removed from the binary (the pixel values for the component are inverted). We used Ensemble of RUS Boosted Decision trees for our classification. Section 5 goes into how to train the classifier.

1. **Probabilistic Recombination: Combining across several bins to generate the final binarized image**

The connected component analysis performed at section 3 and 4 removes nearly all the non-text and fragmented components of text regions from the Delta binary images. At this stage most of the binary images are blank with no connected components left and there are only some images left with non-fragmented text components, spread across different bin sizes and few images having fragments of the background or text. In our experiments we observed bin sizes closer to to have the maximum probability of being a text component. We denote as the *probability that a data pixel in the delta image of is a text pixel.*

We next create a *probability image* () which gives us a probability value

for each pixel. The dimensions are the same as the input image and each

point in has a value. The higher the value, the higher the probability

of that pixel being a text pixel.

denotes a multiplication of the scalar with each

pixels of matrix to generate a new matrix

The final output( is a binary image which we get by bi-level thresholding

For the threshold is set at half the median of the values in the

1. **Training the classifier**

For training the classifier we need pixel level annotations which will be used to label the connected components obtained from the bins as text or others. Pixel level annotations are provided by Kumar et al.[?] for

1. Street View Text 2010 – 647 Images
2. ICDAR 2003 – 1110 Images
3. ICDAR 2011 Born Digital Images - 918 Images
4. ICDAR 2011 Scene Text Images – 716 Images

We split the dataset and keep half the datasets for training and the other half for Testing and evaluation with existing techniques. We take the images for training

and create the Delta images first as described in section 1. Next, from a pixel

level ground truth(GT) we consider every connected component and find the largest component in each delta image that overlaps with the GT connected component. We calculate the (IOU) of the GT connected component and the connected component from the delta image. If IOU > 0.7 we consider the connected component from the delta image as a text component otherwise it’s a non-text component. We extract features from this delta image component and label its corresponding feature vector as 1 if it is text otherwise label it 0. All other components in the delta image which do not overlap with any GT connected component are labelled as 0. We use the feature vectors and their labels as output to train a classifier. From each database we get around 200,000 labelled feature vectors with around 7% to 8% of data as class 1 and rest are class 0. We use an Ensemble of RUS Boosted Decision trees for our classification with number of learners set to 30

1. **Testing and Evaluation**

After training the classifier we evaluate our system. We compare our model with several state-of-art binarization techniques. Note that our objective is to show the strength of the processing technique we proposed. From the evaluation tables below we see our method competing reasonably well with existing techniques despite using a simple classifier model. The high recall goes onto show text regions are captured very accurately as a connected component with geometric properties preserved.

INPUT IMAGE GROUNDTRUTH OUTPUT(OURS) INPUT IMAGE GRUNDTRUTH OUTPUT(OURS)



Fig 5. Some sample output from our system.

**RESULTS ON ICDAR 2003**

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Precision | Recall | F-SCORE |
| Otsu [?] | 0.86 | 0.90 | 0.87 |
| Kittler[?] | 0.75 | 0.89 | 0.78 |
| Sauvola [?] | 0.68 | 0.87 | 0.74 |
| Niblack [?] | 0.65 | 0.83 | 0.67 |
| Wolf [?] | 0.81 | 0.91 | 0.84 |
| Kasar [?] | 0.72 | 0.64 | 0.65 |
| Milyaev [?] | 0.71 | 0.69 | 0.63 |
| Howe [?] | 0.76 | 0.84 | 0.76 |
| Bilateral[?] | 0.84 | 0.85 | 0.83 |
| Mishra[?] | 0.82 | 0.91 | 0.86 |
| **OURS** | 0.78 | 0.93 | 0.85 |

**RESULTS ON ICDAR 2011**

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Precision | Recall | F-SCORE |
| Otsu [?] | 0.87 | 0.91 | 0.88 |
| Kittler[?] | 0.79 | 0.89 | 0.80 |
| Sauvola [?] | 0.75 | 0.86 | 0.79 |
| Wolf [?] | 0.73 | 0.81 | 0.71 |
| Niblack [?] | 0.83 | 0.90 | 0.71 |
| Kasar [?] | 0.65 | 0.47 | 0.85 |
| Milyaev [?] | 0.72 | 0.73 | 0.63 |
| Howe [?] | 0.76 | 0.87 | 0.76 |
| Bilateral[?] | 0.89 | 0.87 | 0.83 |
| Mishra[?] | 0.86 | 0.91 | 0.88 |
| **OURS** | 0.81 | 0.94 | 0.87 |

**RESULTS ON Street View Text**

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Precision | Recall | F-SCORE |
| Otsu [?] | 0.64 | 0.83 | 0.70 |
| Kittler[?] | 0.55 | 0.81 | 0.62 |
| Sauvola [?] | 0.52 | 0.78 | 0.60 |
| Wolf [?] | 0.52 | 0.76 | 0.57 |
| Niblack [?] | 0.58 | 0.81 | 0.66 |
| Kasar [?] | 0.70 | 0.71 | 0.69 |
| Milyaev [?] | 0.52 | 0.66 | 0.51 |
| Howe [?] | 0.62 | 0.77 | 0.64 |
| Bilateral[?] | 0.64 | 0.79 | 0.68 |
| Mishra[?] | 0.64 | 0.82 | 0.71 |
| **OURS** | 0.56 | 0.90 | 0.69 |

**RESULTS ON BDI dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Precision | Recall | F-SCORE |
| Otsu [?] | 0.77 | 0.92 | 0.83 |
| Kittler[?] | 0.57 | 0.88 | 0.63 |
| Sauvola [?] | 0.54 | 0.94 | 0.75 |
| Niblack [?] | 0.59 | 0.94 | 0.71 |
| Kasar [?] | 0.55 | 0.65 | 0.58 |
| Milyaev [?] | 0.48 | 0.68 | 0.61 |
| Howe [?] | 0.43 | 0.93 | 0.52 |
| Bilateral[?] | 0.75 | 0.86 | 0.79 |
| Mishra[?] | 0.70 | 0.90 | 0.80 |
| **OURS** | 0.71 | 0.95 | 0.81 |