Document binarization using topological clustering guided Laplacian Energy Segmentation

Kalyan Ram Ayyalasomayajula
Dept. of Information Technology
Uppsala University
Uppsala, Sweden
kalyan.ram@it.uu.se

Anders Brun
Dept. of Information Technology
Uppsala University
Uppsala, Sweden
anders.brun@it.uu.se

***Abstract*—The current approach for text binarization proposes a clustering algorithm as a preprocessing stage to an energy-based segmentation method. It uses a clustering algorithm to obtain a coarse estimate of the background (BG) and foreground (FG) pixels. These estimates are used as a prior for the source and sink points of a graph cut implementation, which is used to efficiently find the minimum energy solution of an objective function to separate the BG and FG. The binary image thus obtained is used to refine the edge map that guides the graph cut algorithm. A final binary image is obtained by once again performing the graph cut guided by the refined edges on a Laplacian of the image.**

***Keywords*-Image Processing; Classification; Machine Learning; Graph-theoretic methods.**

I. Introduction

Binarized images record each pixel as BG or FG by preserving most of the visual information in the image. A high-quality binarization significantly simplifies and aids in further computation to be performed on the image. The challenges commonly faced in this area are related to uneven illumination of the image; bleed through; fading or paling of the ink in some areas; smudges, stains and blots covering the text; text on textured BG and handwritten documents with heavy-feeble pen strokes for cursive or calligraphic effects to name a few make the task of binarization very subjective. In particular most cited achievements in this field include the work of Otsu [1], Niblack[2], and Sauvola et. al.[3] try to tackle these issues, however no universal solution exists.

Recent work by Howe [4] attempts to employ the use of four simple steps to solve majority of the above stated issues. Firstly, the Laplacian of the image is taken to measure the divergence of the intensity gradient, achieving illumination invariance. Secondly, the cues obtained from the Laplacian operator are aggregated across the entire image to determine plausible source and sink estimates. Thirdly, the results of Canny edge detection guides the the binarization procedure to favour discontinuities in binarization only when they coincide with detected edges. Finally the global fitness function is solved with graph cut (maximum flow) methods, with the source-sink cues provided to efficiently compute the optimal binarization. The second and third steps can be repeated once again with refined edge estimates to better the binarization output. This algorithm was placed near the top on both the DIBCO-09 and H-DIBCO-11 assessments. The current work shows that the results can be further improved through better source-sinks cues and edge estimates that are essential for the performance of the algorithm.

Clustering methods have been used in the past as a means to have preliminary estimates about the data in semi-supervised and machine learning algorithms. In particular we find two of the clustering algorithms namely Ordering points to identify the clustering structure (OPTICS) [5] and Mean Shift Clustering (MSC) [6] of interest here. The former has proven useful in clustering high dimensional data [5] and recursive use of the later was proved to be helpful in image binarization [7]. The current work develops a theoretical framework which enables use of the two algorithms interchangeably. An attempt has been made to device a clustering approach drawing upon the benefit of structure and simple threshold limits that OPTICS has to offer combined with the speed and ease of computation from MSC. The output from the clustering algorithm is used in the source-sink estimates and to refine the edge map essential for Howe's algorithm.

II. Binarization Procedure

A. Motivation

The Howe's binarization algorithm can be split into two stages the first stage gives a binarized image by performing a graph cut on a divergence map of the input image guided through edges from a Canny edge detector with high threshold (t_{hi}) as depicted in the oval outline of Fig.01. This binarized image is used in the second stage

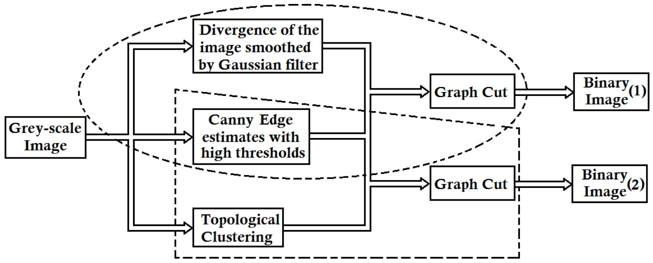
Grey-scale Image
Divergence of the image smoothed by Gaussian filter
Canny Edge estimates with high thresholds
Topological Clustering
Graph Cut
Graph Cut
Binary Image(1)
Binary Image(2)

Figure 1. Comparison of first stage between the Howe's (oval outline) and current approach (trapezoidal outline).

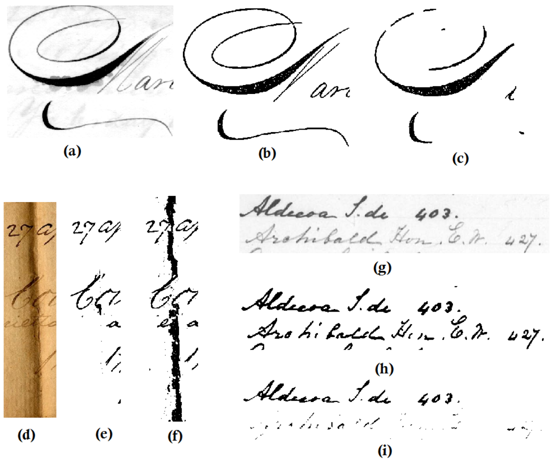

Figure 2. Advantages of Howe's algorithm over clustering. (a),(d),(g) are parts from the original image, (b),(e),(h) are from Howe's output and (c),(f),(i) are from clustering output.

to refine the edges from a Canny edge detector with lower threshold (t_{lo}) by keeping the edges in the FG and neglecting those in BG, followed by performing a graph cut on the divergence map with these refined edges. The advantage of this procedure is better delineation of edges as depicted in Fig.02 (a)-(c); separation of BG edges that over lap in FG as depicted in Fig.02 (d)-(f); and conservation of heavy and faint stokes in FG Fig.02 (g)-(i). On the other hand when only the topological clustering to be described in the following section, if used it performs a coarse separation of BG and FG with a threshold on cluster size (t_{sz}). The advantage of this procedure is better BG estimation Fig.03 (a)-(c), (d)-(f); and better FG estimation (g)-(i), (j)-(l). We would like to preserve the best of these two approaches in the current algorithm, where the first stage in Howe's algorithm (oval outline of Fig.01) is replaced by the cluster approach (trapezoidal outline of Fig.01) keeping the second stage intact. The theoretical framework for the clustering algorithms is explained in the next section, however insight behind the clustering algorithm is explained in subsection on clustering algorithm.

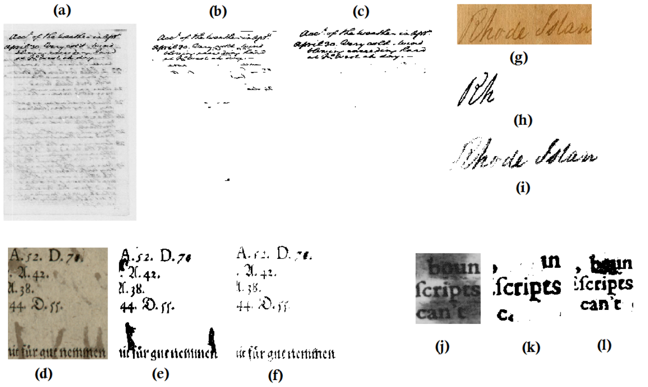
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)

Figure 3. Advantages of clustering over Howe's approach. (a),(d),(g), (j) are parts from the original image, (b),(e),(h),(k) are from Howe's output and (c),(f),(i),(l) are from clustering output.

B. Clustering Method

1) Theoretical Framework: For the clustering algorithm we choose the following three feature for each pixel $i)$ intensity value for each pixel (I), $ii)$ gradient in x direction (dI_x) and $iii)$ the gradient in y direction (dI_y). Each pixel can then be represented as a point in a three dimensional space by the ordered triplet (I, dI_x, dI_y) we define this as the *binarization space* denoted by $\mathscr{B}$. It can be noted that the region $\mathscr{B}$ is bounded within the region

$$\mathscr{S} = \begin{cases} 0 \leq I \leq 255, \\ -255 \leq dI_x \leq 255, \\ -255 \leq dI_y \leq 255. \end{cases} \tag{1}$$

We explicitly define the following definitions borrowed from the works of Ankerst.et.al [5] in OPTICS as some definition will be slightly altered for mathematical convenience.

Def: Given an $\varepsilon > 0$ the set of all points that are within the Euclidean distance of ε from a a point Q is defined as *ε-neighbourhood* of a point Q denoted as $N_\varepsilon(Q)$.

Def: The key idea of density-based clustering is to find set of points exceeding a certain number within a given region. More formally, given a $N_\varepsilon(Q)$ we are interested in all

$$N_\varepsilon(Q) \geq M_n$$

where M_n is the threshold on the number of points, such a point Q is called the *core object* (CO).

Def: Given a core object Q, point P is *directly density-reachable* (DDR) from point Q wrt. (ε, M_n) in a set of points $\mathscr{D}$ if $P \in N_\varepsilon(Q)$.

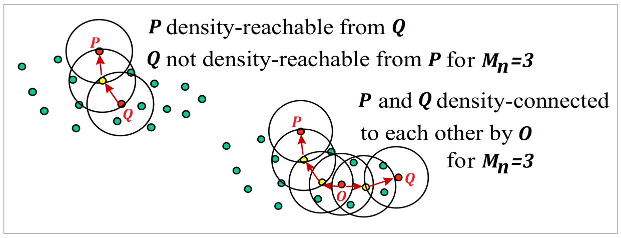
P density-reachable from Q
Q not density-reachable from P for M_n=3
P
Q
P and Q density-connected to each other by O for M_n=3
P
O
Q

Figure 4. Figure adapted from [5] explaining density reachability and connectivity.

Def: A point P is *density-reachable* (DR) from an point Q wrt. (ε, M_n) in the set of points $\mathscr{D}$ if there is a chain of points $Q = P_1, \dots, P_n = P$; $P_i \in \mathscr{D}$ $\;s.t\;$ P_{i+1} is DDR from P_i.

Def: An point P is *density-connected* (DC) from an point Q wrt. (ε, M_n) in the set of points $\mathscr{D}$ if $\exists\ O \in \mathscr{D}$ $\;s.t\;$ P, Q are DR from O.

Def: Let $\mathscr{D}$ be a set of points. A *cluster* wrt. (ε, M_n) denoted by $Cl(\varepsilon, M_n)$ in $\mathscr{D}$ is a non-empty subset of $\mathscr{D}$ satisfying the following conditions:

$i)$ Maximality: $\forall P, Q \in D$: if $P \in Cl(\varepsilon, M_n)$ and Q is DR from P wrt. (ε, M_n), then also $Q \in Cl(\varepsilon, M_n)$.
$ii)$ Connectivity: $\forall P, Q \in Cl(\varepsilon, M_n)$: P is DC to Q wrt. $(\varepsilon, M_n) \in \mathscr{D}$.

Def: Every point not contained in any cluster is *noise* wrt. (ε, M_n).

Though the following argument is in general valid for any set of point we restrict ourselves to $\mathscr{B}$. Given a set of points $\mathscr{D}$ in the binarization space $\mathscr{B}$ we make the following observations.
i. $\mathscr{D}$ is a bounded set, as $\mathscr{D} \subset \mathscr{B} \subset \mathscr{S}$ and $\mathscr{S}$ is bounded.
ii. We define $E = \{d(P_i, P_j) : P_i, P_j \in \mathscr{D}$ where $d(P_i, P_j)$ is Euclidean-distance between $P_i, P_j\}$, E is bounded as the the underlying set $\mathscr{D}$ is bounded. Let $\varepsilon_m, \varepsilon_M$ denote the infimum and supremum of E respectively and we take some fixed ε_0 $\;s.t\;\varepsilon_0 < \varepsilon_m$ and denote $\mathscr{E} = E \cup \{\varepsilon_0\}$.
iii. Similarly the set $\mathscr{D}$ is a finite set of points we denote $M_N = card(\mathscr{D})$, where $card(A)$ denotes the cardinality of the set A and denote $\mathscr{M} = \{1, 2, 3, \dots, M_N\}$.

Given points $P, Q, R \in \mathscr{D}$, are core objects wrt. some $\varepsilon \in \mathscr{E}$ and $M_n \in \mathscr{M}$, let $Cl_A(\varepsilon, M_n)$ denote the cluster relation with respect to the core object $A \in \mathscr{D}$.
i. $Cl_A(\varepsilon, M_n)$ is not reflexive, *i.e.* P may not form a core object for all (ε, M_n) combinations.
ii. $Cl_A(\varepsilon, M_n)$ is symmetric, *i.e.* $P \in Cl_Q(\varepsilon, M_n) \Rightarrow Q \in Cl_P(\varepsilon, M_n), M_n \geq 2$.
iii. $Cl_A(\varepsilon, M_n)$ is transitive, *i.e.* $P \in Cl_Q(\varepsilon, M_n), Q \in Cl_R(\varepsilon, M_n) \Rightarrow P \in Cl_R(\varepsilon, M_n), M_n \geq 3$.

However if we consider the noise wrt. (ε, M_n) as the cluster $Cl_\phi(\varepsilon, M_n)$ and define every point $\in Cl_\phi(\varepsilon, M_n)$ as a core object then the relation $Cl_A(\varepsilon, M_n)$ becomes an equivalence relation.

2) Theorem: Given a set $\mathscr{D}$ and $\varepsilon \in \mathscr{E}$ and $M_n \in \mathscr{M}$, $\mathscr{D}$ forms a topology under the following collections

a) $C_{Cl} = \bigcup\limits_{\lambda \in \Lambda} Cl_A(\varepsilon, M_n)$; the resulting topology denoted by τ_{Cl}

b) $C_N = \bigcup\limits_{\lambda \in \Lambda} N_\varepsilon(A)$; the resulting topology denoted by τ_N, where $\bigcup\limits_{\lambda \in \Lambda}$ denotes the arbitrary union of all these sets.

Proof:
We take the Neighbourhoods definition of topological spaces, we need to verify the following axioms for each of the two cases. Unless otherwise stated, $\varepsilon_i \in \mathscr{E}$ and $M_j \in \mathscr{M}$, $\forall\ i, j \in \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers.

a. For the C_{Cl} collection

- *Each point belongs to every one of its neighbourhoods.*

  $Cl_A(\varepsilon, M_n)$ are essentially maximal $N_\varepsilon(A)$ subject to the condition that they contain at least M_n points (neighbours) so $P \in Cl_A(\varepsilon, M_n)$ for which the maximal and M_n-neighbours count is satisfied.

- *Every superset of a neighbourhood of a point P in $\mathscr{D}$ is again a neighbourhood of P.*

  We see that the equivalence classes wrt. $(\varepsilon_1, M_n), (\varepsilon_2, M_m)$ say $[Cl_A(\varepsilon_1, M_n)]$ and $[Cl_B(\varepsilon_2, M_m)]$ respectively define a partition on $\mathscr{D}$, say $[Cl_A(\varepsilon_1, M_n)] \subseteq [Cl_B(\varepsilon_2, M_m)]$ and $P \in [Cl_A(\varepsilon_1, M_n)] \Rightarrow P \in [Cl_B(\varepsilon_2, M_m)]$ as any partition of a set is mutually exclusive and exhaustive.

- *The intersection of two neighbourhoods of P is a neighbourhood of P.*

  We need to consider two cases here,

  Case 1:
  The point P forms clusters with $Cl_Q(\varepsilon_1, M_n), Cl_R(\varepsilon_2, M_m)$ then as $Cl_A(\varepsilon, M_n)$ is an equivalence relation these clusters belong to

equivalence classes $[Cl_Q(\varepsilon_1, M_n)]$ and $[Cl_R(\varepsilon_2, M_m)]$ respectively.

Both these equivalence classes wrt. (ε_1, M_n) and (ε_2, M_m) define their respective partitions and every point belonging to both these clusters such as P will belong to $[Cl_Q(\varepsilon_1, M_n)]$ and $[Cl_R(\varepsilon_2, M_m)]$.

Hence one such class must be contained in another by the maximality property of clusters, say $Cl_Q(\varepsilon_1, M_n) \subset Cl_R(\varepsilon_2, M_m) \Rightarrow Cl_Q(\varepsilon_1, M_n) \cap Cl_R(\varepsilon_2, M_m) = Cl_Q(\varepsilon_1, M_n)$ which is a neighbourhood of P.

Case 2:
The point P forms a cluster with $Cl_Q(\varepsilon_1, M_n)$ and noise wrt. $Cl_R(\varepsilon_2, M_m)$.

Then $P \in Cl_Q(\varepsilon_1, M_n)$ and $P \in Cl_\phi(\varepsilon_2, M_m)$ and $Cl_Q(\varepsilon_1, M_n) \cap Cl_\phi(\varepsilon_2, M_m) = Cl_Q(\varepsilon_1, M_n) \setminus Cl_R(\varepsilon_2, M_m) \subseteq Cl_\phi(\varepsilon_2, M_m)$ (by definition of $Cl_\phi(\varepsilon_2, M_m)$).

Hence is a neighbourhood of P.

- *Any neighbourhood N of P contains a neighbourhood M of P such that N is a neighbourhood of each point of M.*
  Let $P \in Cl_A(\varepsilon, M_n)$ as per our construction of the neighbourhoods $P \in Cl_P(\varepsilon_0, 1)$. And $Cl_P(\varepsilon_0, 1) \subset Cl_A(\varepsilon, M_n)$ which are both neighbourhoods of P and this is true $\forall P \in \mathscr{D}$.

Hence $\mathscr{D}$ is a topology with the collection C_{Cl} we denote this topology by τ_{Cl}.

b. For the C_N collection

We make a few observations on the set $\mathscr{E}$ before we verify the following axioms. $\mathscr{E} \subset \mathbb{R}$ ($\mathbb{R}$ is the set of all real numbers), $\mathbb{R}$ being an ordered field we can arrange elements in $\mathscr{E}$ in the sequence $\varepsilon_0, \varepsilon_1 (= \varepsilon_m), \varepsilon_2, \ldots, \varepsilon_M$ *s.t* $\varepsilon_i < \varepsilon_j, \forall$ i < j.

- *Each point belongs to every one of its neighbourhoods.*

- *Every superset of a neighbourhood of a point P in $\mathscr{D}$ is again a neighbourhood of P.*

  By definition of the ε-neighbourhood of a point P the above axioms are satisfied.

- *The intersection of two neighbourhoods of P is a neighbourhood of P.*
- *Any neighbourhood N of P contains a neighbourhood M of P such that N is a neighbourhood of each point of M.*

  As per our construction $P \in N_{\varepsilon_0}(P)$, hence the above two axioms are true $\forall P \in \mathscr{D}$.

Hence $\mathscr{D}$ is a topology with the collection C_N we denote this topology by τ_N.
□

*3) Theorem: Given the topologies τ_{Cl} and τ_N on $\mathscr{D}$; $\tau_{Cl} \subset \tau_N$*i.e. *the topology τ_N is finer (stronger or larger) than the topology τ_{Cl}*

Proof:
By definition $Cl_A(\varepsilon, M_n)$ are essentially maximal $N_\varepsilon(A)$ subject to the condition that $card(N_\varepsilon(A)) \geq M_n$ neighbours. Hence, $C_{Cl} \subset C_N \Rightarrow \tau_{Cl} \subset \tau_N$. □

C. Clustering Algorithm

We spend some time in this section on showing how the discussion on OPTICS and MSC so far are related and their interchangeable operability. The OPTICS framework enables us to view any given collection of points as a large cluster with sub-clusters of varying densities within it. This approach is useful in the database classification where the entire collections of data-points can be viewed as a dendrogram with dense sub-clusters embedded within a bigger diffused clusters. This approach classifies clusters based on *reachability-metric* values [5], that can be viewed as a function mapping each point in $\mathscr{B}$ to a real number depending on proximity of the point to a core object. The clusters are identified if the variation between the reachability-metric values of two successive ordered points exceed a certain threshold [5]. This mechanism of identifying clusters from the local minima in the reachability metric plot works for the binarization or image segmentation cases, but can prove to be an over kill. The MSC on the other hand employs radially symmetric kernels operating on ε-neighbourhoods to detect dense sub-clusters within a larger diffused cluster; however by iteratively reducing the bandwidth of these kernels we can reconstruct the whole

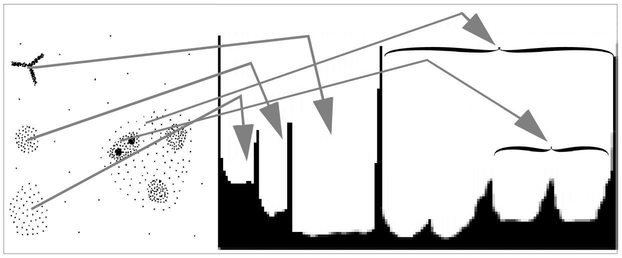

Figure 5. Typical reachability-plots adapted from [5] for a data set with hierarchical clusters of different sizes, densities and shapes. Each local minima corresponds to a cluster as indicated in the plot.

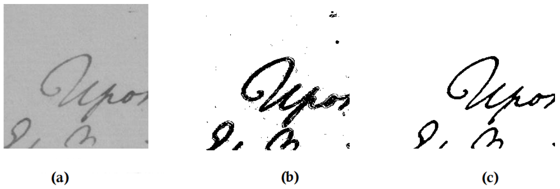

Figure 6. Typical output image blocks where (a) is the original image (b) is the output from OPTICS (c) is the output from MSC.

structure of the data-points. This fact can be verified by observing that OPTICS operates on τ_{Cl} and MSC operates on τ_N. Since we have proved that $\tau_{Cl} \subset \tau_N$ an algorithm such as MSC can capture all the classification aspects that OPTICS has to offer and OPTICS can capture all the cluster information with the difference in their operation being how the noise gets separated; hence they are inter-changeable for the classification of BG and FG. For coarse separation of the BG from FG we threshold on the size of all the dense clusters put together that were separated from the larger diffused cluster, which is equivalent to detecting the local minima following the peaks in the reachability metric in OPTICS depicted in Fig.05. This enables us to deploy MSC to detect clusters as predicted by OPTICS but at a faster convergence rate. Fig.06 shows a typical 128x128 block of the output for OPTICS and MSC when run on *2009_H01.bmp* test image.

Algorithm 1 Algorithm for Clustering guided binarization

```
[1] Bandwidth = $BW_{large}$
[2] DataPoints = [ I, dIx, dIy ]
[3] Divergence = Divergence(image)
[4] EdgeImage = Canny(image, $t_{hi}$)
[5] Clusters = MeanShiftCluster(DataPoints, Bandwidth)
while ( sizeof(maximal(Clusters)) $\geq t_{sz}$) do
    [6] Bandwidth = Bandwidth / 2
    [7] Clusters = MeanShiftCluster(DataPoints, Bandwidth)
end while
[8] BinaryImage = [ maximal(Clusters)== 0 ]
[9] EdgeImage = Compute8neighbours(BinaryImage, EdgeImage)
[10] BinaryImage = GraphCut(Divergence, EdgeImage)
[11] WeakEdgeImage = Canny(image, $t_{lo}$)
[12] EdgeImage = EdgeImage | (WeakEdgeImage & BinaryImage)
[13] BinaryImage = GraphCut(Divergence, EdgeImage)
```

III. Experiments

The experiments were conducted in MATLAB so we adhere to the MATLAB convention for data representation.

Table I
Comparison of the results for FM, p-FM, PSNR, DRD

File Name	FMeasure		p-FMeasure		PSNR	
	Howe	MSC	Howe	MSC	Howe	MSC
2013_HW05.bmp	69.40	85.72	71.24	88.55	15.95	20.17
2012_H13.png	63.05	78.09	66.50	81.33	15.58	17.24
2011_PR6.png	85.81	92.04	86.66	93.01	18.13	20.94
2011_PR2.png	74.98	79.91	77.21	82.46	11.46	12.70
2011_HW1.png	85.41	73.79	87.80	75.53	13.91	10.72
2009_H05.bmp	86.05	75.35	87.25	76.25	19.80	16.71
2013_HW03.bmp	85.59	77.96	88.85	81.14	17.49	15.98
Mean	87.20	87.70	89.67	90.17	17.70	17.86
Median	88.70	90.11	91.53	91.88	18.14	18.29
Mode	92.00	91.00	95.00	96.00	19.00	19.00
Avg. Gain/Loss	0.56		0.57		0.18	

Table II
Comparison of the results for DRD, Recall, Precision

File Name	DRD		Recall		Precision	
	Howe	MSC	Howe	MSC	Howe	MSC
2013_HW05.bmp	19.17	6.10	92.85	92.87	55.40	79.60
2012_H13.png	8.03	5.16	48.84	69.61	88.90	88.92
2011_PR6.png	8.17	3.32	93.77	93.78	79.10	90.36
2011_PR2.png	13.45	9.60	91.32	91.30	63.60	71.05
2011_HW1.png	8.38	19.07	93.80	94.07	78.40	60.70
2009_H05.bmp	4.27	11.28	84.69	85.44	87.46	67.39
2013_HW03.bmp	3.73	5.65	80.21	67.60	91.75	92.07
Mean	4.45	4.20	87.73	88.31	87.94	88.05
Median	3.42	3.16	90.92	91.21	89.97	89.78
Mode	3.00	3.00	91.00	91.00	92.00	93.00
Avg. Gain/Loss	-0.29		0.59		0.22	

For each image (M rows, N columns) we compute the x-derivative, y-derivative of the image and create an (3xMN) matrix of the image by augmenting the linearised matrices of image pixel values, and its x and y derivatives as shown in STEP[2] of the algorithm. The current clustering approach is an unsupervised clustering method and we exploit the fact that variation in intensities or their derivative will not exceed 255 so the MSC bandwidth is set to a value of 128 or 256 as indicated in STEP[1]. The MSC algorithm is run on the data-points matrix with reducing bandwidth parameters. The resulting cluster sizes are monitored until the size threshold on the FG clusters (t_{sz} = 2% of total image size) is met as indicated in STEP[4]-[7]. The binarized image is generated by marking BG pixels in black and FG in white (STEP [8]) as it will be used to preserve all the edges (STEP[4] with thresholds t_{hi} indicated in Howe's algorithm [4]) with at least one pixel classified as FG by saving all the 8-connected neighbours of FG pixels along the edges (STEP [9]). The graph-cut was done on the binarized image acting as source-sink prior guided by the modified edges with their respective weights ($W_{ss} = 0.66$, $W_{edge} = 25.0$). The resulting binary image was used to refine the edges with lower thresholds (STEP[11] with thresholds t_{lo} indicated in Howe's algorithm [4]). A second graph-cut on the divergence map of the image (STEP[3]) with the refined edges (STEP[12]) yields the final binary image.

The $DIBCO$ [8] datasets form 2009 to 2013 consisting of 66 images were used in testing. This dataset consists of a varied collection with samples from hand written and machine written documents with cursive/calligraphic text; images with textured / varying intensity BG; images with bleed-through; documents with stains/smudges or folds that tend to match FG when digitally replicated; and images with a low contrast between the FG and BG due to colour, varying pressure in hand strokes or fading. These images from a rigours if not a comprehensive test set for measuring the performance of the classifier. The experiments were run on a 64-bit Intel i7-4600U processor with 16 GB RAM, the tests would run for about a minute for most of the images to about an hour for files with low contrast such as (*2012_H10.png; 2012_H11.png*). The performance of the clustering algorithm can be improved by reusing the clusters data obtained through the coarser bandwidth. The parameter used for the clusters sizes were used from the estimates gained through the reachability plots of OPTICS and the size threshold were reused with MSC clustering. The threshold W_{ss}, W_{edge} used on the clustered output ever estimated by normalizing the W_{ss}, W_{edge} estimates used in the divergence map in the ratio of the source-sink average grey-scale values. The following tables present the results for the files where a gain or loss of more than 5% as reported from the $DIBCO$-evaluation tool was observed. The top four rows report the files where the algorithm gained and the next three rows show where the algorithm lost wrt. Howe's method. The mean, median, mode values (*mode figures are rounded to the nearest integers values to allow repetition*) and average gain over all the 66 test cases have been reported in the bottom four rows. The file names have been prefixed with the corresponding year of the dataset release for disambiguity. The evaluation metrics used here are standard metrics employed in evaluating imaging algorithms [8].

IV. Conclusion

The current work aims at presenting a simple unsupervised classifier that can predict the background and foreground in an image accurately. A theoretical framework behind the working of the classifier was developed. The classification approach combines the structural advantage OPTICS has to offer with the computational speed of Mean Shift Clustering. The result from the experiments show a considerable gain in the case where the documents suffer from bleed through and textured backgrounds. The classifier is also efficient in revealing the text behind smudged/blotted foreground. The current approach however suffer in increased noise due to misclassification of background with gradation in intensities and high cursive text where some edge information is not propagated by the classifier. The current approach can be further refined through better tuning of the weights parameter used for the graph-cut as opposed to a simple heuristic estimate based on average intensity ratios. The method can also gain to estimate edge pixel better from employing features such as distance transform to respect edge pixels.

Acknowledgment

The authors would like to thank Fredrik Wahlberg and Tomas Wilkinson of Dept. of Information Tech., Uppsala University for their constructive criticism in improving the manuscript; "Vetenskapsrådet grant 2012-5743" and the "q2b – From Quill 2 Bytes initiative at Uppsala University".

References

[1] N. Otsu. *A threshold selection method from grey level histogram.* IEEE Trans. on System, Man, Cybernetics, 19(1):62-66, January 1978.

[2] W. Niblack. *An Introduction to Digital Image Processing*. PrenticeHall, Englewood Cliffs, New Jersey, 1986.

[3] N. Sauvola and M. Pietikainen. *Adaptive document image binarization*. Pattern Recognition, 33(2):225-236, January 2000.

[4] N.R. Howe. *A Laplacian Energy for Document Binarization*. International Conference on Document Analysis and Recognition, 2011.

[5] M. Ankerst, M. M. Breunig, H. Kriegel and J. Sander. *OPTICS: Ordering Points To Identify the Clustering Structure*. ACM Press, 49–60, 1999.

[6] D. Comaniciu, P. Meer. *Mean shift: A robust approach toward feature space analysis*. IEEE Trans. Pattern Anal. Machine Intell., 24:603-619, 2002.

[7] R. Rodrguez. *Binarization of medical images based on the recursive application of mean shift filtering : Another algorithm*. Advances and Applications in Bioinformatics and Chemistry 2008: 1 1-12.

[8] Pratikakis, I., Gatos, B., Ntirogiannis, K., *ICDAR 2011 document image binarization contest (DIBCO 2011)*. Proc. Internat. Conf. Document Anal.Recognition, pp. 15061510, 2011.