DEVELOPING DOCUMENT IMAGE RETRIEVAL SYSTEM

Konstantinos Zagoris
Image Processing and Multimedia Laboratory
Department of Electrical & Computer Engineering
Democritus University of Thrace, 67100 Xanthi, Greece

Kavallieratou Ergina
Department of Information and Communication Systems Engineering
University of the Aegean, Samos 83100, Greece

Nikos Papamarkos
Image Processing and Multimedia Laboratory
Department of Electrical & Computer Engineering
Democritus University of Thrace, 67100 Xanthi, Greece

ABSTRACT

A system was developed able to retrieve specific documents from a document collection. In this system the query is given in text by the user and then transformed into image. Appropriate features were in order to capture the general shape of the query, and ignore details due to noise or different fonts. In order to demonstrate the effectiveness of our system, we used a collection of noisy documents and we compared our results with those of a commercial OCR package.

KEYWORDS

Document Retrieval, Word Spotting, Segmentation, Information Retrieval

1. INTRODUCTION

Huge quantities of document images are created and stored in image archives without having any indexing information. In order to satisfactorily exploit these collections of document images, it is necessary to develop effective techniques to retrieve the document images.

OCR packages were applied to documents in order to convert them to text. Thus, Edwards et al (2004) described an approach to transcribing and retrieving Medieval Latin manuscripts with generalized Hidden Markov Models. More recently, with the improvement in document image processing (DIP) field, techniques that make use of images instead of OCR were also introduced. Leydier et al (2005) uses DIP techniques to create a pattern dictionary of each document and then he performs word spotting by selecting the feature of the gradient angle and a matching algorithm. Konidaris et al. (2007) proposes a technique for keyword guided word spotting in historical printed documents. He creates synthetic image words as query and performs word segmentation using dynamic parameters and hybrid feature extraction. Rath et al (2004) presents a method for retrieving large collections of handwritten historical documents using statistical models.

In this paper, we propose a Document Image Retrieval System (DIRS). The proposed technique encounters the document retrieval problem using a word matching procedure. This technique performs the word matching directly in the document images bypassing OCR and using word-images as queries. The entire system consists of the Offline and the Online procedures. In the Offline procedure, the document images are analyzed and the results are stored in a database. Three main stages, the preprocessing, the word segmentation and the feature extraction stages, constitute the offline procedure. A set of features, capable of capturing the word shape and discard detailed differences due to noise or font differences are used for the

word-matching process. The Online procedure consists of four components: the creation of the query image, the preprocessing stage, the feature extraction stage, and finally, the matching procedure.

The section 2 describes the overall system while the section 3 presents the features that are used. The matching procedure is described in detail in Section 4. Sections 5 and 6 describe the implementation of the proposed system and present some test results, respectively. Finally, in Section 7 we draw some conclusions and the future directions of our research.

2. THE DOCUMENT IMAGE RETRIEVAL SYSTEM (DIRS)

The overall structure of the proposed system is presented in Figure 1. It is constituted of two different parts: the Offline and the Online procedure.

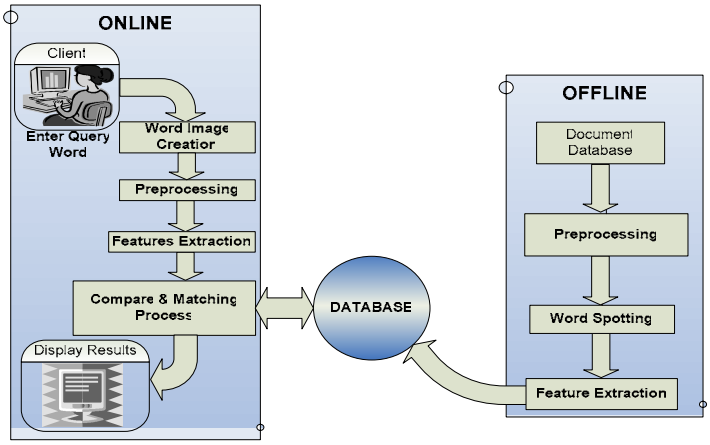
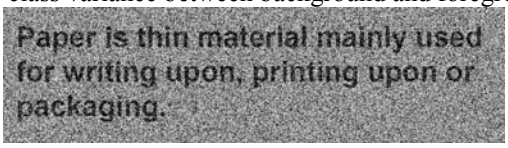


Figure 1. The overall structure of the Document Image Retrieval System

2.1 Preprocessing Stage

In the Offline operation, the document images are analyzed in order to localize the word limits and the results are stored in a database. This procedure consists of three main stages. Initially, the document images pass the preprocessing stage which consists of a Median 5x5 filter (Figure 2(a)), in order to face the existence of noise e.g in case of historical or badly maintained documents, and a binarization method (Figure 2(b)). The Median filtering is a nonlinear signal processing technique developed by Tukey (1974) that is useful for noise suppression in images (Pratt 2007). The binarization is achieved by using the well known Otsu technique (Otsu 1979). This technique performs binarization through the histogram of the image by minimizing the inter-class variance between background and foreground pixels.

a. 

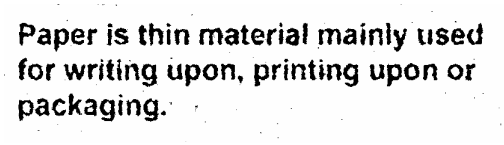
b. 

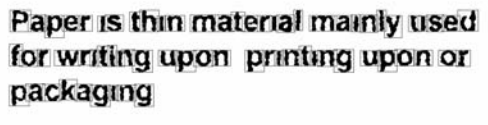
Figure 2. The Preprocessing Stage: (a) The original noisy document (b) After applying the median 5x5 filter and the Otsu technique.

2.2 Word Segmentation

The word segmentation stage follows the preprocessing stage. Its primary goal is to detect the word limits. This is accomplished by using the Connected Components Labeling and Filtering method.

After indentifying all the Connected Components (CCs), the most common height of the document CCs (CCch) is calculated. Since the noise of some documents can change significantly the shape of the extracted word images, the CCs with height less than 70% of the CCch are rejected. In [8] has been proven that the height of a word can reach the double of a character mean size due to presence of ascenders and descenders. This means that the applied CCs filtering can only reject areas of punctuation points, noise.

The next step includes the expansion of the left and right sides of the resulted CCs (Figure 3(a)) by 20% of the CCch as depicts the Figure 3(a). Finally, to locate the words the overlapping CCs are merged (Figure 3(b)). Due to the facts that Kavallieratou et al (2002) presents about the mean character size and having filtering out accents, noise and punctuation marks, it is rather rare character of the same word to have distance greater than 20% of the CCch and different words to be closer than that.

a. 

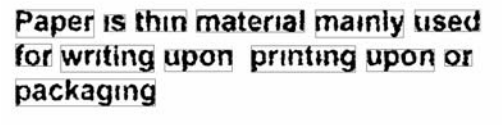
b. 

Figure 3. (a) The expanded CCs (b) The final extracted words after the merging of the expanded CCs

3. FEATURES

The proposed system is based on six powerful features that are extracted from every word capable of capturing the word similarities and discarding the small differences due to remained noise or different style of fonts. The six-feature-set is:

Width to Height Ratio: The width to height ratio of the word forms important information concerning the word shape.

Word Area Density: This feature represents the percentage of the black pixels included in the word-bounding box. It is calculated by using the following relation:

$$E = 100\frac{(BP)}{(IW)\cdot(IH)} \tag{1}$$

where (BP) is the number of black pixels in the word bounding box, (IW) is the width and (IH) is the height of the word bounding box in pixels.

Center of Gravity: It represents the Euclidean distance from the word's center of gravity to the upper left corner of the bounding box. In order to calculate this, the vertical and horizontal center of gravity must be determined by the following equations:

$$C_x = \frac{M_{(1,0)}}{M_{(0,0)}} \quad (2), \qquad C_y = \frac{M_{(0,1)}}{M_{(0,0)}} \tag{3}$$

Where C_x is the horizontal center and C_y the vertical center of gravity and $M_{(p,q)}$) the geometrical moments of rank $p+q$:

$$M_{pq} = \sum_x \sum_y \left(\frac{x}{width}\right)^p \left(\frac{y}{height}\right)^q f(x,y) \tag{4}$$

The x and y determine the image pixels. The image is binary, so the $f(x,y)$ is considered to be 1 when the pixel (x,y) is black and 0 when the pixel is white. The division of the x and y by the width and the height of the image, respectively, cause the geometrical moments to be normalized and be invariant of the size of the word. Finally, the Center of the Gravity feature is defined as equal to the Euclidean distance from the upper left corner of the image:

$$COG = \sqrt{C_x^2 + C_y^2} \tag{5}$$

Vertical Projection: This feature consists of a vector with twenty (20) elements, extracted from the smoothed and normalized vertical projection of the word image (Figure 4). These elements correspond to the first twenty (20) coefficients of the Discrete Cosine Transform (DCT) of the smoothed and normalized

vertical projection. The smooth and normalized vertical projection has the average width and height properties for all the word images and its calculation consists of the following steps:

Step 1: From the equation (6) a new projection $VP[i]$ is produced which has width l and maximum height h. In our proposed method, the l and h are equal to the mean width and height respectively of all the word bounding boxes. For our experiment database is: $l = 167$ and $h = 50$.

$$VP[i] = \frac{h}{h_{\max}} VP_{orig}\left[i \cdot \frac{l_{orig}}{l} \right] \tag{6}$$

The l_{orig} is the original width of the original projection $VP_{orig}[i]$ and $h_{\max}$ the height of the word bounding box.

Step 2: The final normalized vertical projection depicted in the Figure 4(c) is created after applying a $5x1$ mean mask to the projection $VP[i]$. This way, the final projection is more robust to the changes of the size and type of fonts.

a.

b.

c.

Figure 4. A visual representation of the Vertical Projection calculation: (a) Original Image (b) The Vertical Projection of the Original Image (c) The smoothed and normalized Vertical Projection

a.

b.

c.

d.

e.

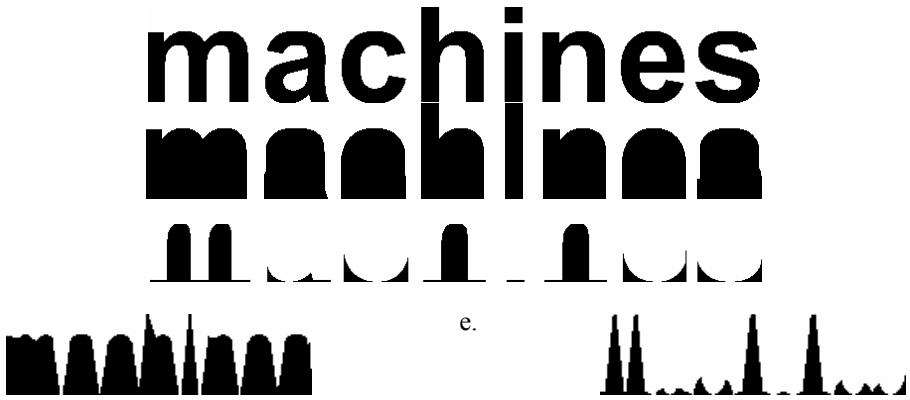


Figure 5. A visual representation of the Top-Bottom Shape Projection calculations: (a) Original Image (b) Original Top Shape Projection (c) Original Bottom Shape Projection (d) The smoothed and normalized Top Shape Projection (e) The smoothed and normalized Bottom Shape Projection

Top – Bottom Shape Projections: As it is shown in Figure 5, the Top–Bottom Shape Projections can be considered as signatures of the word shape. These signatures lead to a feature vector of 50 elements, where the first 25 values are the first 25 coefficients of the smoothed and normalized Top Shape Projection DCT (Figure 5(d)) and the rest 25 values are equal to the first 25 coefficients of the smoothed and normalized Bottom Shape Projection DCT (Figure 5(e)).

In order to calculate the Top Shape Projection, the word image is scanned from top to bottom. As it is shown in Figure 5(b), the first time a black pixel is found all the following pixels of the same column are converted to black.

The Bottom Shape Projection is found similarly. As it is shown in Figure 5(c), the word image is scanned from bottom to top and all the pixels are converted to black until a black pixel is found. The smoothed and

normalized Shape Projections (Figure 5(d)-(e)) are calculated as described at the Vertical Projection paragraph.

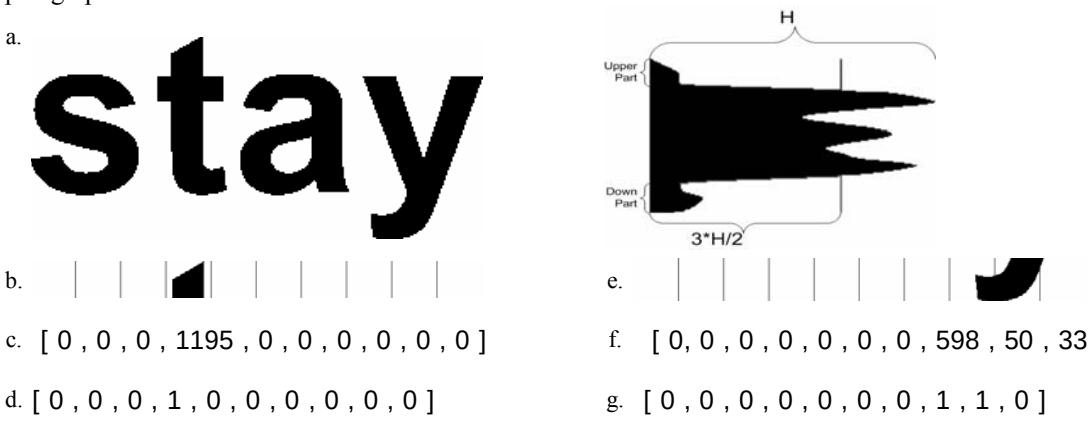


Figure 6. A visual representation of the Upper Grid and Down Grid Information feature extraction procedures: (a) The original word-box image and its horizontal projection (b) The extracted upper part and its separation in 10 parts (c) The number of black pixels in each part (d) The final vector of UGF (e) The extracted lower part and its separation in 10 parts (f) The number of black pixels in each part (g) The final vector of DGF

Upper Grid Features: The Upper Grid Features (UGF) is a ten element vector with binary values which are extracted from the upper part of each word image. In order to calculate this, initially the image's horizontal projection is extracted, and from it, the upper part of the word is determined by the following algorithm:

Step 1: Smooth the horizontal projection with a mean 5×1 mask.

Step 2: Starting from the top, find the position i in the horizontal projection histogram $V(i)$ where $V(i) \geq \frac{3}{2} \cdot H$ as depicted in Figure 6(a). The H is the maximum height of the horizontal projection ($\max\{V(i)\}$). If the position i is below the half of the world then the word has no upper part.

Step 3: Find the position $k \in (0,i)$ in the V(i) horizontal projection histogram when $V(k) - V(k-1) \leq 0$ .Then the k defines the upper part of the word. If the k has very small value (3 or 2) the word has no upper part.

Then the upper part of the word is separated in ten same parts as depicted in Figure 6(b). The number of black pixels is counted for each part. If this number is bigger than the height of the extracted image, the relative value of the vector is set to 1 otherwise it is set to 0. Figures 6(b)-(c) illustrates an example. The height of the extracted image is 43. The obtained feature vector is shown in Figure 6(d).

Down Grid Features: As the name suggests, they are similar to the UGF but they are extracted from the lower part of the word image. The Down Grid Features (DGF) are calculated by using the method of the UGF extraction but this time the search is starting from the bottom of the $V(i)$ horizontal projection histogram. The output is again a ten element vector with binary values. Figures 6(e)-(f) give an example of the DGF extraction. This time the height of the extracted image, in our experimental set, was 50 pixels. Figure 6(g) shows the final feature vector.

4. COMPARISON

Every time the user enters a query word, the proposed system creates an image of the query word with font height equal to the average height of all the word-boxes obtained through the Word Segmentation stage of the Offline operation. In the implemented DIRS for our experimental set the average height is 50. The font type of the query image is Arial. However, the smoothing and normalizing of the various features described before, suppress small differences between various types of fonts.

Next, the created query image is being processed in exactly the same way as the document word images. That includes the application of Preprocessing and Feature Extraction procedures.

The matching procedure can identify the word images of the documents that are more similar to the query word through the extracted feature vectors. Figure 7 illustrates the comparison technique.

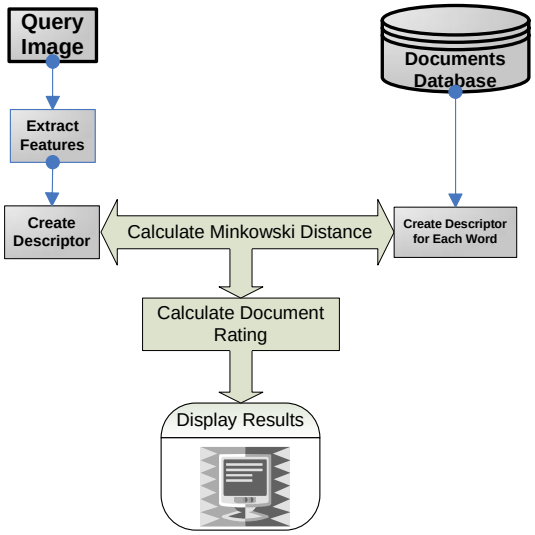


Figure 7. The matching process

First, a descriptor is created by the 7 extracted features. The first element is the Weight to Height feature, the second the Image Area Density feature and the third the Center of Gravity feature. The following twenty (20) elements are the ones extracted from the Vertical Projection feature and the next fifty (50) from the Top – Bottom Shape Projection features. Finally, the last twenty (20) elements are the ones extracted from the Upper and Down Grid features divided by 10 in order to prevent to overpower the others features. The rest features values are normalized from 0 to 1.

Next, the Minkowski L1 distance is obtained between the descriptor of the query image word and the descriptor of each word in the database:

$$MD(i) = \sum_{k=1}^{93} \left| Q(k) - W(k,i) \right| \tag{7}$$

where $MD(i)$ is the Minkowski distance of the i word, $Q(k)$ is the query descriptor and $W(k,i)$ is the descriptor of the i word.

Then the similarity rate of the remaining words is computed. The rate is a normalized value between 0 and 100 which depicts how similar are the words of the database with the query word. The similarity rate for each word is defined as:

$$R_i = 100\left(1 - \frac{MD(i)}{\max(MD)}\right) \tag{8}$$

where R_i is the rate value of the word i, $MD(i)$ is the Minkowski distance of the i word and $\max(MD)$ the maximum Minkowski distance found in the document database.

Finally, the system presents the documents that contain the words in descending order with respect to the corresponding rate. In our implementation, the documents presented to the user are those that have a similarity rate above 70.

5. IMPLEMENTATION

The proposed system is implemented with the help of the Visual Studio 2008 and is based on the Microsoft .NET Framework 3.5. The programming language which is used is the C#.

The image documents which are included in the database are created artificially from various texts and then noise was added in order to implement in parallel a text search engine which makes easier the verification and evaluation of the search results of the DIRS system. Furthermore, the database which is being used by the implemented DIRS is the Microsoft SQL Server 2005.

The web address of the implemented system is the http://orpheus.ee.duth.gr/irs2_5.

6. EVALUATION

The Precision and the Recall metrics have been used to evaluate the performance of the proposed system. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. In our evaluation, the precision and recall values are expressed in percentage.

The evaluation of the proposed system was based on 100 document images. In order to calculate the Precision and Recall values thirty searches have been made using random words. Table 1 shows those random words. The Precision and Recall values obtained are depicted in Figure 8. The Mean Precision and the Mean Recall values for this experiment are 87.8% and 99.26% respectively.

Table 1. The thirty words which are used in the evaluation

1. details	2. potential	3.religion	4. technology	5. advent
6. smoothing	7. culture	8. world	9. between	10. further
11. number	12. Greek	13. might	14. century	15. homage
16. period	17. taxes	18. living	19. growth	20. churches
21. neural	22. foreign	23. smaller	24. extensively	25. eventually
26. diplomatic	27. demands	28. political	29. region	30. break

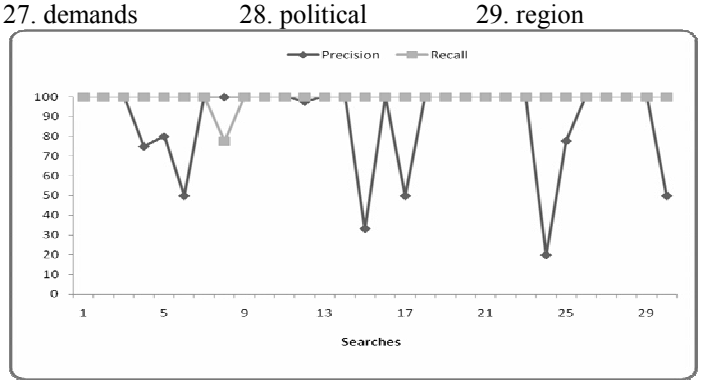


Figure 8. The variation of the Precision and Recall coefficients of the proposed DIRS for 30 searches. The Mean Precision is 87.8% and the Mean Recall is 99.26%

The overall time for the above mentioned thirty searches in our AMD Athlon 64 4200+ testing server with 2 GB of RAM is 11.53 seconds while the mean time for each search is approximately 0.38 seconds.

Furthermore, the same 100 document images were scanned from the FineReader® 9.0 (2007) OCR program and the results are searched for the same thirty words of Table 1. The Precision and Recall values obtained are depicted in Figure 9(a). The Mean Precision and the Mean Recall values are 76.667% and 58.421%, respectively, lower enough than the proposed system.

In order to test the robustness of the features relative to the type of fonts, the query font was changed to "Tahoma" and the same thirty searches have been made (Table 1). The Precision and Recall values obtained are depicted in Figure 9(b). The Mean Precision and the Mean Recall values are 89.44% and 88.05%, respectively.

a. b.

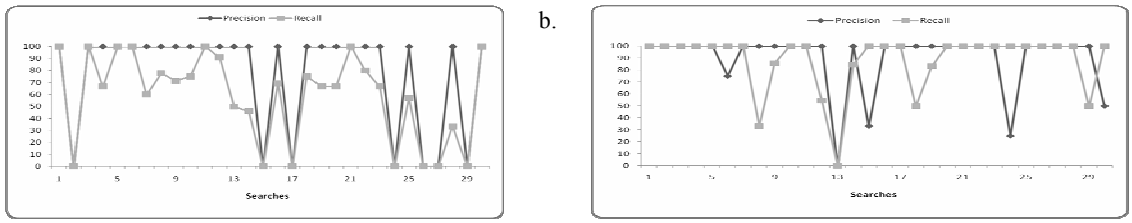


Figure 9. (a) The variation of the Precision and Recall coefficients of the FineReader® 9.0 OCR program for 30 searches. The Mean Precision is 76.67% and the Mean Recall is 58.42% (b) The variation of the Precision and Recall coefficients for 30 searches with the query font name "Tahoma". The Mean Precision is 89.44% and the Mean Recall is 88.05%

7. CONCLUSION

A document retrieval system was presented here. The proposed system makes use of document image processing techniques, in order to extract powerful features for the description of the word images. Seven meaningful features were used, namely the Weight to Height ratio, the Image Area Density, the Center of Gravity feature, twenty DCT coefficients extracted from the Vertical Projection, fifty DCT coefficients extracted from the Top – Bottom Shape Projection and twenty elements extracted from the Upper and Down Grid. These features was selected in such way that describe satisfactorily the shape of the query words while at the same moment they suppress small differences due to noise, size and type of fonts.

Our experiments were performed on a collection of noisy documents and gave Mean Precision 87.8% and Mean Recall 99.26%. The same experiment performed in the same database for a commercial OCR package gave lower results while experiments for different size and style of fonts didn't produce significant change to the performance.

ACKNOWLEDGEMENT

This work is co-funded by the project "PENED 2003-03EΔ679".

REFERENCES

ABBYY FineReader®, 2007. Available: http://finereader.abbyy.com/, Accessed December 2007.

Edwards, J. et al, 2004. Making latin manuscripts searchable using gHMM's. *Proc. 18th Annual Conf. on Neural Information Processing System.* Cambridge, USA, pp. 385-392.

Kavallieratou, E. et al, 2002. A*n Off-line Unconstrained Handwriting Recognition System.* International Journal of Document Analysis and Recognition, No 4, pp. 226-242.

Konidaris, T. et al, 2007. Keyword-Guided Word Spotting in Historical Printed Documents using Synthetic Data and User Feedback. *International Journal on Document Analysis and Recognition*, Vol. 9, No 2, pp. 167-177.

Leydier, Y. et al, 2005.Textual Indexation of Ancient Documents. *DocEng'05.* Bristol, UK, pp. 111- 117.

Otsu, N., 1979. *A threshold selection method from gray-level histograms.* IEEE Trans. Systems, Man, and Cybernetics, Vol 9, pp. 62-66.

Pratt, W. K., 2007. *Digital image processing, Fourth Edition*. John Wiley & Sons, Inc., New York, NY.

Rath, T.M. et al. A Search Engine for Historical Manuscript Images. *Proc. ACMSIGIR conference.* Sheffield, UK, pp.369–376.

Tukey, J.W., 1974. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.