International Journal of Technology (2017) 5: 947-956
ISSN 2086-9614 

# HISTOGRAM EQUALIZATION IMPLEMENTATION IN THE PREPROCESSING PHASE ON OPTICAL CHARACTER RECOGNITION

Peter Pangestu[1*], Dennis Gunawan[1], Seng Hansun[1]

[1]*Computer Science Study Program, Faculty of Engineering and Informatics*
*Universitas Multimedia Nusantara, Jl. Scientia Boulevard, Gading Serpong, Tangerang, Banten 15811, Indonesia*

(Received: September 2016 / Revised: February 2017 / Accepted: October 2017)

## ABSTRACT

A 2014 report from Digital Marketing Philippines stated that the number of web applications with visual content as their main product has increased significantly. Image processing technology has also undergone significant growth. One example of this is optical character recognition (OCR), which can convert the text on an image to plain text. However, a problem occurs when the image has low contrast and low exposure, which potentially results in information being hidden in the image. To address this problem, histogram equalization is used to enhance the image's contrast so the hidden information can be shown. Similar to X-ray scanning used in the medical field, histogram equalization processes scanned images that have low brightness and low contrast. In this study, histogram equalization was successfully implemented using OCR preprocessing. The test was done with a dataset that contains dark background images with low light text; the successful outcome resulted in the ability to show 74.95% of the information hidden in the image.

*Keywords:* Contrast enhancement; Histogram equalization; Image processing; Information hiding; Optical character recognition

## 1. INTRODUCTION

In 2014, Digital Marketing Philippines, a digital consultant, issued an analysis of a survey conducted on visuals used in web applications. The analysis explained that the use of visual content in web applications has developed very rapidly (Digitalmarketingphilippines.com, 2014). This suggests that visual message delivery methods, such as infographics and charts, can provide users with more information and draw greater attention. These developments have also been followed by the development of image processing technologies, such as face detection, object detection, and optical character recognition (OCR) (MathWorks.com, 2016).

Using OCR, images can be produced through a process that includes a series of arrangements, background separation, and matching characters (Sánchez et al., 2012). OCR can be achieved using machines that are currently popular. An OCR readeris not versatile; it must be supported by good conditions and it must meet suitable image matching criteria. Light-dark settings and resolution also affect the performance of an OCR reader. The success of machine translation is also affected by the engine as well as the techniques used in the OCR (Abbyy-developers.eu, 2015). Therefore, OCR cannot be run optimally if the image inserted does not support the desired conditions. Thus, preprocessing is very important in order to create an image that is

*Corresponding author's email: peterpangestu@live.com, Tel: +62-21-5422-0808, Fax: +62-21-5422-0800
Permalink/DOI: https://doi.org/10.14716/ijtech.v8i5.877

ready to be processed. One important preprocessing step is the improvement of the image with dark and light colors.

No OCR reader can read all the conditions of an image perfectly (Abbyy-developers.eu, 2015). Various tests have been conducted with a variety of datasets (iapr-tc11.org, 2015). However, the majority of the images contained in the dataset collection have had fairly good image conditions (quite bright and good contrast), and they have supported the delivery of clear information from the image (no hidden information). Therefore, to ensure that OCR can be used for a variety of image conditions, improvements in the color, light, and dark elements used in an image (color adjustment) are needed. Some of the methods commonly used to improve the condition of an image include histogram equalization, Wiener filtering, median filtering, decorrelation stretch, and unsharp mask filters (MathWorks.com, 2016).

In the present study, histogram equalization was selected as the image enhancement method because it is similar to X-ray scanning that is used in the medical field to scan organs. Histogram equalization was used to clarify the background and the object of regional differences, and to identify information hidden in the images due to low light and low contrast (Akhlis & Sugiyanto, 2011). In addition, this method is considered fairly common and easy to apply to an image (Alginahi, 2010).

## 2. THEORETICAL BASIS

In this section, we explain the theoretical basis that informed the present research study, including ocr, color adjustment, image histogram, and histogram equalization.

### 2.1. Optical Character Recognition (OCR)

OCR allows a computer to recognize the characters in an image (Mithe et al., 2013). OCR consists of a scanner to read the text and software to process the image. OCR is no longer fixated on the physical form of the machine; rather it has been widely applied to desktop computers and even smartphones.

According to Alginahi's (2010) scientific work, "Preprocessing Techniques in Character Recognition", in general, translation is done through several stages. The first task is done by adjusting the distribution of the color composition, for example light-dark images, contrast, and noise reduction. The second task entails establishing the setting thresholding in which the colors at each pixel become only black (RGB # 000000) or white (RGB # FFFFFF). The third task, the segmentation process, is carried out to separate the region between the background and text. In the fourth task, the separation process is initiated to separate the character segments. In the fifth task, typography recognition is applied to determine the characteristics of the font to be used. In the final task, each of the characters of the pattern comparators available in the storage media is matched. However, the present study focused on the arrangement of the colors listed in the first stage of processing.

Zybert (2014) posted an article on the Nedocs developer forum entitled, "How Does Optical Character Recognition Work?" He noted that, prior to being text ready, the characters of an image will go through a series of processes before being recognized one-by-one (Zybert, 2014).

Each of these processes is explained below (Zybert, 2014):

*1. Preprocessing*

Preprocessing is the initial stage of translation. In this stage, noise removal is done so that the patterns that are not needed can be omitted.

*2. Segmentation*

Segmentation is the stage where the application determines the location of the text in the image

in order to facilitate the selection of the scanning area. This process divides the image into two regions: background and text areas. After performing the division, OCR will only do further processing on the text that has been segmented in a given region.

*3. Normalization*
In this stage, the shape, pattern, or thickness of each character is detected. If OCR accepts input in the form of a word composed of various sizes, the normalization process will detect these differences in order to form a uniformcharacters' size, making it easier to perform OCR extraction stage thereafter.

*4. Extraction*
In this stage, the hallmark of the character is detected and normalized. Consequently, the OCR can know the typeface of the text that is used.

*5. Recognition*
Recognition is the final stage of the translation of the information generated in the extraction step. Characters that have been separated one-by-one will then be compared with the patterns of characters corresponding to the existing pattern, which are usually stored in a database.

## 2.2. Color Adjustment

This section describes some of the terms that are commonly discussed in image processing.

*1. Gray Value*
Gray values (GV) refers to the value of measuring the brightness of a point in the entire region of the image displayed in gray scale. The region with a bright color has a high GV, while regions with dark colors have low GV. Figure 1 shows a graph with GV dominant in the middle (the image tends to be gray). Figure 2 shows image contrasts in which the GV is relatively even. This causes the image to have a deployment range of darkness.

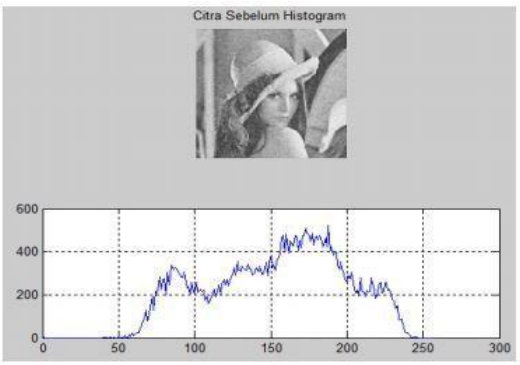


Figure 1 Image with a middle-dominant GV (Ahmad & Hadinegoro, 2012)

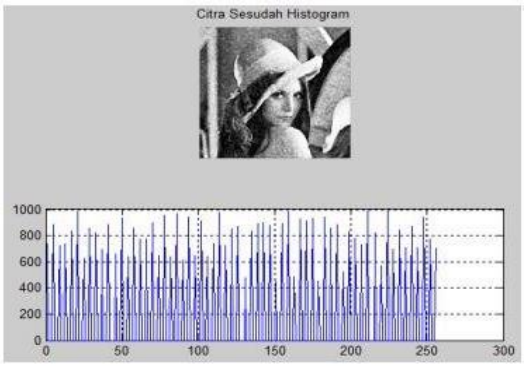


Figure 2 Image with an equalized GV (Ahmad & Hadinegoro, 2012)

*2. Highlights and Shadows*
The elements presented in Figure 1 and Figure 2 represent the darkness in an image. The highlights are the light given to the object; in other words, the part that receives a lot of light. The shadows are the image area that is either not affected or only slightly gets light so as to produce a dark field. The area between points one and two is the so-called shadow, while the area between points two and three is the highlights.

*3. Brightness and Contrast*
Brightness is the amount of exposure in an image (Xcitex, 2010). By increasing the brightness value, the GV of the image will be increased by the specified value. Contrast is the light and dark variation settings in an image, so that differences between the background and the object

region can be identified (Xcitex, 2010). The contrast can be increased by decreasing the GV of the shadow area and increasing the GV of the highlights, creating a more noticeable difference between the light and dark regions.

### 2.3. Image Histogram

An image histogram (IH) is a collection of the statistical data of an image that stores information about the distribution of the GV (Krutsch & Tenorio, 2011). The dataset will form a graph. An image dominated by dark colors has a greater GV spread in the beginning and less of a GV spread at the end, while the bright-dominated region has a solid distribution in the final GV (Figure 3).

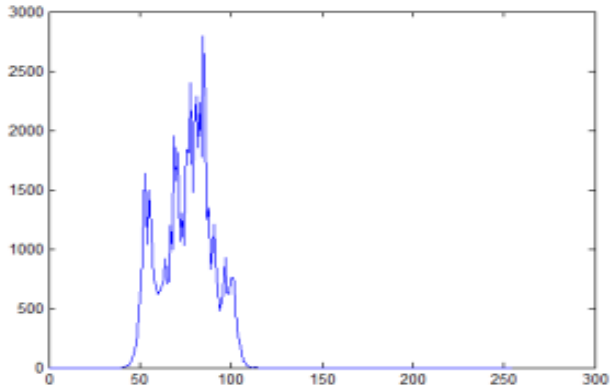


Figure 3 Image histogram (Krutsch & Tenorio, 2011)

### 2.4. Histogram Equalization

Histogram equalization is used in image processing to strengthen the color and increase the contrast (Gonzalez & Woods, 2008). This method is applied by distributing the GV level. In general, this is done by increasing the lower limit of the range of colors to the darkest point, and by decreasing the upper limit of the range of colors to the brightest point. Currently, histogram equalization is popularly used in the medical field, especially to improve organ scanning images (Akhlis & Sugiyanto, 2011). In distribution, narrowing this range will change the cumulative frequency of the deployment GV so as to form a linear line on a graph IH that satisfies the equation $X = Y$. This causes the image to have more contrast than the original image. This method was chosen because it is easy to implement and it improves the contrast in an image (Ahmad & Hadinegoro, 2012). Based on Rachman (2014), the following formula was used to calculate the GV using histogram equalization:

$$S = \sum_{j=0}^{k} \frac{n_j}{n} \quad (1)$$

where $n$ is the number of data, $j$ is the range of GV, and $n_j$ is the number of data on the GV.

Figure 4 Histogram before applying histogram equalization

Figure 5 Histogram after applying histogram equalization

## 3. RESEARCH METHODOLOGY

The research methodology used in the present study was implemented in the following stages.

### 3.1. Learning and Consulting

A literature review was conducted to identify previous research on this topic. This phase was done by collecting supporting data associated with the present research study. The data collection was done by reviewing various types of scientific work, such as books, journals, and articles. This phase is carried out so that the research study can be conducted in accordance with the provisions presented in previous studies in order to produce a valid conclusion.

### 3.2. Designing the Application and Identifying the Analysis Requirements

After collecting the supporting data, we conducted a needs analysis to determine the standard to be used in research. In addition to the analysis, we designed the application that will be used as media in the study. This design resulted in several documents, namely flowcharts and the structure of the storage table. Figure 6 shows the procedures we undertook to implement the histogram equalization method used in this research.

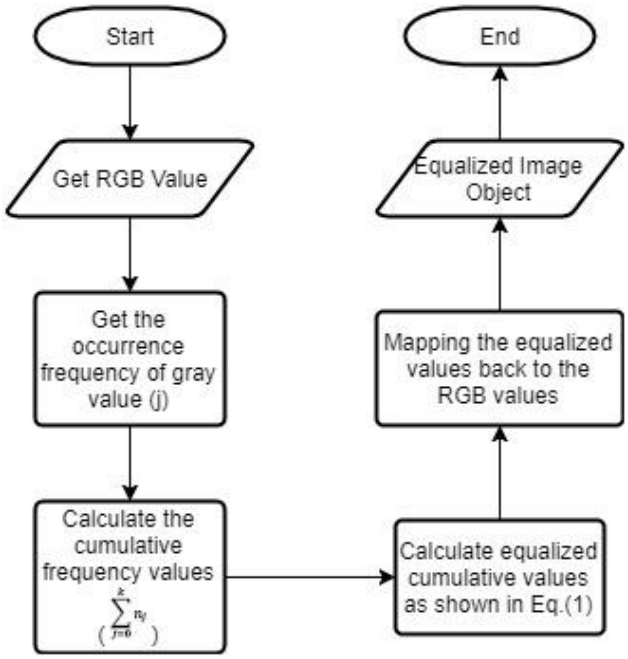


Figure 6 Flowchart of the histogram equalization method used in the study

### 3.3. Programming

After establishing the design and determining the analysis requirements, we programmed the application. The application served as a media liaison between the users and the systems used for implementation. The application was designed so users can enter their own samples and obtain rapid analysis of them.

### 3.4. Testing and Debugging

After completing the programming stage, the application was tested. The tests were conducted using all the functions that were made in the programming stage (Stage 3). In addition to testing all the functions, we also performed tests on all the possibilities that could occur in the application when it is used by a user.

### 3.5. Collecting Samples

After the testing was completed, the application was deemed ready to be used in real-time by users. Therefore, we collected the samples. The sample collection was done by looking for a

random sample in accordance with the needs generated in the designing the application and identifying the analysis requirements (Stage 2).

### 3.6. Analysis of the Test Results

After obtaining a variety of samples from users, we further analyzed the data to determine the impact of the implementation results and conclusions from the application of the histogram equalization method used in the study.

### 3.7. Report and Documentation

To complete the research activities, we wrote a report to document the procedures and findings of the research study.

## 4. RESULTS AND DISCUSSION

After designing and implementing the program, further testing was done using the samples that were collected. The samples collected by searching for dark background samples tended to be black, have low contrast, and were not crowded. The analysis aimed to determine if the image conditions were in accordance with the application of histogram equalization in general. The tests were performed offline using a sample dataset, which contained images with a dark background picture in dominant black (RGB # 000000). Then, some texts were added on each image. Selection of the color of the text was done by adding a hexadecimal value 10 for each component of red, green, and blue (RGB) of the background; for example, RGB #000000 to RGB #101010. The test results are presented in Table 1.

Table 1 The testing results

| No. | Title | Text | | Target | Success (%) | |
|---|---|---|---|---|---|---|
| | | Before histogram equalization | After histogram equalization | | Before histogram equalization | After histogram equalization |
| 1 | sugar.png | | lele jumbo | lele jumbo | 0 | 100 |
| 2 | manus.png | | I will kill you | I will kill you | 0 | 100 |
| 3 | phantom-assassin-dota-2-dota-2740(1).png | | assasin | assasin | 0 | 100 |
| 4 | cool_dark_wallpaper.png | | cool | cool | 0 | 100 |
| 5 | 12June2012-Low-light-focusing-lrg | | LION KING | LION KING | 0 | 100 |
| 6 | images_(5).jpg | | the dragon | the dragon | 0 | 100 |
| 7 | images.jpg | | S N THE DARKEST PLACES | THERE IS LIGHT EVEN IN THE DARKEST PLACES | 0 | 48.6486486 |
| 8 | Normal21.bmp | POLITICS H appuintment he | | POLITICS H Appointment he general manage | 63.1578947 | 0 |
| 9 | Normal22.bmp | TAIW | | TAIW Taipei | 40 | 0 |
| 10 | Normal23.bmp | Not to | | not to | 100 | 0 |
| 11 | Normal25.bmp | Chand Said | | e arguments change said | 47.3684211 | 0 |

| No. | Title | Text | | Target | Success (%) | |
|---|---|---|---|---|---|---|
| | | Before histogram equalization | After histogram equalization | | Before histogram equalization | After histogram equalization |
| 12 | Norhe mal26.bmp | the World | | the world | 100 | 0 |
| 13 | Normal27.bmp | mm mm mbigs she | | news feature abies lishe | 20.8333333 | 0 |
| 14 | Normal28.bmp | atlc exerci of detention tlso reneged | | atlc exerci of detention lse reneged | 97.14286 | 0 |
| 15 | Normal29.bmp | | | in | 0 | 0 |
| 16 | Shadow1.bmp | THURSDAY | zrea ysterda | THURSDAY | 100 | 0 |
| 17 | Shadow2.bmp | massacre and mothers tell it like it | tell it like | a massacre and mothers tell it like it is | 93.93939 | 30.3030303 |
| 18 | Shadow4.bmp | eather tai9i lhudershw hsinch9 th | | Weather Taipei thundershowe hsinchu thundershow | 62.7907 | 0 |
| 17 | low-key_cat.jpg | | | low light cat | 0 | 0 |
| 18 | 7d0GbFV.png | | ON YOUR MIND | ON YOUR MIND | 0 | 100 |
| 19 | 610.png | DOGGY | DOGGY | DOGGY | 100 | 100 |
| 20 | Flower-1.jpg | | | flower | 0 | 0 |
| 21 | black_hat.jpg | | a man with black hat | a man with black hat | 0 | 100 |
| 22 | 6808332673_3fe8de3766_b.jpg | yellow flower | yellow flower | yellow flower | 100 | 100 |
| 23 | Black_and_white_photography_tips_DCM120.feature.darktones_lowkey.jpg | | | this is low light photo | 0 | 0 |
| 24 | dark wallpaper 28.png | | dark wallpaper | dark wallpaper | 0 | 100 |
| 25 | dark_wallpaper_a3222.jpg | | kucing | kucing | 0 | 100 |
| 26 | final_500.png | | FINAL 500 | FINAL 500 | 0 | 100 |
| 27 | jf93UfN.png | | pardon my face | pardon my face | 0 | 100 |
| 28 | download (5).jpg | your face | | your face | 100 | 0 |
| 29 | dark_souls_wallpaper_a3243 | | cat face | cat face | 0 | 100 |
| 30 | sony-vaio-wallpaper-black-wallpaper-hd-10c.png | | NEED MOTIVATION | NEED MOTIVATIO N | 0 | 100 |
| Average | | | | | 10 | 74.95 |

Table 1 presents the character recognition results from a database that contains a sample set of pictures with a dark background, which tended to be black (RGB # 000000). As depicted in Table 1, the results show a significant improvement in the sample after histogram equalization in comparison to before histogram equalization. An average of 10% of the characters can be read on a sample that was not equalized, and an average of 74.95% of the characters can be read on a sample after histogram equalization was implemented.

The calculation of the success of OCR on each image is only determined based on the number of character extraction results in comparison to the targets set by the human eye. This is because OCR cannot produce 100% accuracy for characters' recognition; but depends solely on the approach or technique applied by the OCR. Therefore, the target setting is applied manually through the vision of the human eye (Holley, 2009). We applied the formula used at The 4th Annual Test of OCR Accuracy, presented by Rice et al. (1995), which is described as follows:

$$\frac{character\,count - character\,error}{character\,count} \times 100\% \quad (2)$$


Figure 7 Image that has not been equalized (source: images (5).jpg)

Figure 7 shows one of the samples contained in the dataset (data no.6). The image shows an object resembling a dragon with a dominant black background (RGB # 000000). From Table 1 it can be seen that the character recognition results of the image presented in Figure 7 before equalization is 0%.

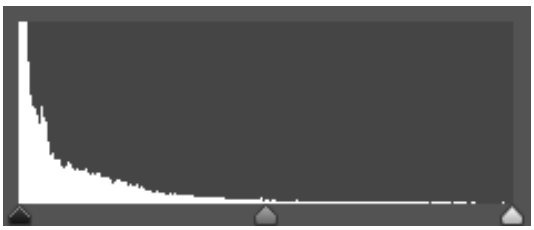

Figure 8 Histogram for "images_(5).jpg" before histogram equalization

Figure 8 shows a histogram of 'images_(5).jpg' contained in the dataset. The histogram in Figure 8 shows that the majority of the GV is in the shadow area, and it tends to be dominant in one place. This shows that 'images_(5).jpg' is an image with low light and low contrast characteristics.


Figure 9 Image that has been equalized (source: images_(5).jpg)

Figure 9 shows the results of histogram equalization for the image presented in Figure 7. Figure 9 shows an increase in the contrast between the background and the object, which is demonstrated in Figure 10. This affects the rate of the appearance of the text object pasted on the image, i.e., 'the dragon'. The information presented in Table 1 shows that the recognition results of the sample reached 100%.

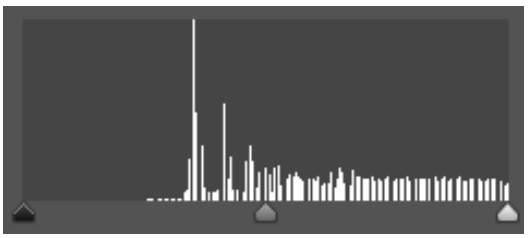

Figure 10 Histogram for “images_(5).jpg” after histogram equalization

## 5. CONCLUSION

Histogram equalization was successfully implemented during the OCR preprocessing phase by using a web-based program (PHP). The research results demonstrate that implementing histogram equalization during OCR preprocessing improved the OCR performance. The dataset that contained a collection of predominantly black and dark images inserted with dark texts increased the percentage of success to approximately 74.95%.

In the future, additional studies can be conducting using other advanced methods to improve the image contrast without causing a lot of noise, such as adaptive histogram equalization and contrast-limited adaptive histogram equalization. We believe that more advance methods could result in better image contrast than is possible using a normal histogram equalization method.

## 6. REFERENCES

Abbyy-developers.eu, 2015. Image Processing and Binarisation for Camera OCR. Available online at https://abbyy.technology/en:features:ocr:cameraocr-preprocessing-binarisation

Ahmad, N., Hadinegoro, A., 2012. Metode Histogram Equalization untuk Perbaikan Citra Digital. *In*: Proceedings of Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2012, Semarang: Indonesia, INFRM, pp. 439–445

Akhlis, I., Sugiyanto, 2011. Implementasi Metode Histogram Equalization untuk Meningkatkan Kualitas Citra Digital. *Jurnal Fisika,* Volume 1(2), pp. 70–74

Alginahi, Y., 2010. *Preprocessing Techniques in Character Recognition*, Character Recognition, Minoru Mori (Ed.), ISBN: 978-953-307-105-3, InTech. Available online at http://cdn.intechopen.com/pdfs-wm/11405.pdf

Digitalmarketingphilippines.com. 2014. *Amazing Facts and Statistics about Visual Web.* Available online at http://digitalmarketingphilippines.com/wp-content/uploads/2014/01/Amazing-Facts-and-Statistics-about-Visual-Web.jpg

Gonzalez, R., Woods, R., 2008. *Digital Image Processing* (3rd ed). New Jersey: Prentice-Hall

Holley, R., 2009. How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. Available online at http://www.dlib.org/dlib/march09/holley/03holley.html

iapr-tc11.org, 2015. Datasets List - TC11. Available online at http://www.iapr-tc11.org/mediawiki/index.php/Datasets_List

Krutsch, R., Tenorio, D., 2011. *Histogram Equalization.* Guadalajara: Freescale Semiconductor Application Note Number AN4318, Rev 0

MathWorks.com, 2016. Image Processing and Computer Vision Examples. Available online at http://www.mathworks.com/examples/product-group/matlab-image-processing-and-computer-vision

Mithe, R., Indalkar, S., Divekar, N., 2013. Optical Character Recognition. *International Journal of Recent Technology and Engineering (IJRTE)*, Volume 2 (1), pp. 72–75

Rachman, E.M.B.P., 2014. Histogram Equalisation. Available online at http://ilmukomputer.org/wp-content/uploads/2014/02/Histogram-Equalisation-Pengolahan-Citra-Digital.odt

Rice, S.V., Jenkins, F.R., Nartker, T.A., 1995. The Fourth Annual Test of OCR Accuracy. Available online at http://www.expervision.com/wp-content/uploads/2012/12/1995.The_Fourth_Annual_Test_of_OCR_Accuracy.pdf

Sánchez, J., Perronnin, F., de Campos, T., 2012. Modeling the Spatial Layout of Images Beyond Spatial Pyramids. *Pattern Recognition Letters*, Volume 33(16), pp. 2216–2223

Xcitex, Inc., 2010. Image Processing: Brightness, Contrast, Gamma, and Exponential/Logarithmic Settings in ProAnalyst. Available online at http://www.xcitex.com/Resource%20Center/ProAnalyst/Application%20Notes/App%20Note%20151%20-%20Image%20Processing%20Brightness,%20Contrast,%20Gamma%20and%20Exponential.pdf

Zybert, C., 2014. How does Optical Character Recognition Work. Available online at http://nedocs.com/how-does-optical-character-recognition-work/