An analysis of absorbing image on the Indonesian text by using color matching

G A Hutagalung\textsuperscript{1}, Tulus\textsuperscript{2}, Iryanto\textsuperscript{2}, Y F A Lubis\textsuperscript{3}, M Khairani\textsuperscript{3} and Suriati\textsuperscript{3}

\textsuperscript{1}Department of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan-20155, Indonesia
\textsuperscript{2}Department of Mathematics, Universitas Sumatera Utara, Medan-20155, Indonesia
\textsuperscript{3}Department of Informatics, Universitas Harapan Medan, Medan-20217, Indonesia

Email: tulus@usu.ac.id

Abstract. The insertion of messages in an image is performed by inserting per character message in some pixels. One way of inserting a message into an image is by inserting the ASCII decimal value of a character to the decimal value of the primary color of the image. Messages that use characters in letters, numbers or symbols, where the use of letters of each word is different in number and frequency of use, as well as the use of letters in various messages within each language. In Indonesian language, the use of the letter A to be the most widely used, and the use of other letters greatly affect the clarity of a message or text presented in the language. This study aims to determine the capacity to absorb the message in Indonesian language from an image and what are the things that affect the difference. The data used in this study consists of several images in JPG or JPEG format can be obtained from the image drawing software or hardware of the image makers at different image sizes. The results of testing on four samples of a color image have been obtained by using an image size of 1200 X 1920.

1. Introduction
Messages are information that can have different levels of privacy. Not all messages are publicly available. Some messages have some levels of privacy that must be maintained from unauthorized persons to know the content of messages intended for the recipient. One of message concealment method is steganography. This method hides the messages in a media either in the form of images, sounds or other media that can hide the message to the wrong recipients [1]. Digital image is one of the most commonly media used in steganography. This medium can be an option in message insertion, where messages are inserted into some pixels of digital images either in one bit or directly in the pixel decimal value. [2]

Digital images have the absorption capacity of messages to be inserted based on several things. Color values of the digital image can be used to be inserted in the images. The differences of color values will affect the number of words or letters of a sentence to be inserted into the image. Messages have different frequencies of letter use in each language used for the messages. For example, the English language has the letter E that is the highest value in use. On the other hand, the letter A is the most widely used letter reached 19.81% in every word that is used in Indonesian. [3] This study aimed to analyze how the absorption of the digital image to the message in Indonesian language. The analysis is performed to the number of words that can be inserted in the digital image and something that influence differences in absorption image of the messages.

2. Previous Work
Some previous researches on image insertion are performed by using different image types, i.e. color image and grayscale image, to produce a better steganography. In these works there are different ways of insertion that are done because of different color values of the selected image type as a sample [3]. The research on the frequency of use in Indonesian language, letters is carried out by taking samples of formal texts from official speeches, newspapers and through stories written in both fiction and nonfiction. In this research, the study yields an average of 6.6 letters in each word in Indonesian texts [4]. The research on steganography used in Arabic has been done using Novel Approach which used a data input to produce Arabic text using algorithm approach which is done after Encrypt Compressed Secret Message (ECSM) [5].
Research on calculating the frequency of using letters of the text in Malay language has also been done based on the average use of letters in some media in both print and online media, where the letter A in Malay language reaches 19% of its use in every word. Malay and Indonesian language have some similarities in the writing of words, but the difference that often occurs in the meaning and pronunciation of the word used [6]. The authors argue that for the use of text in Malay and Indonesian language have similarities. The previous research in the calculation of the frequency of letters in Malay can be used as a reference in the technique of calculating the frequency of the use of letters in Indonesian language. The N-gram method can be used to detect whether the language used in the text is Indonesian, that does not use English, and is better with the speed of search is also quickly completed with [7].

3. The Proposed Method
To obtain the results of the research objectives, the authors conducted research implementation measures or the research design, in outline, the authors carried out research which consists of the following steps:

1. Analysis and sampling, where the sample or object of this research are some digital images having different size and color values consisting of different image types (monochrome, grayscale, color).
2. Introduction of values on each sample image tested, because each image has a different byte value that is tailored to the constituent color of the image.
3. Testing by adjusting the color values in digital images with the value of the occurrence of letters in the Indonesian text.
4. The number of letters that appear adjusted to the average use in the word to find the number of words that can be accommodated in the image.
5. Results and conclusions from testing.

The data used in this study consists of several images in JPG or JPEG format can be obtained from the image drawing software or hardware of the image makers at different image sizes. For a sample image of the author using 9 pieces following image:

Table 1. Sample image

| NUMBER | TYPE    | SIZE       |
|--------|---------|------------|
| 1      | Monochrome | 256 x 256 |
| 2      | Monochrome | 450 x 250 |
| 3      | Monochrome | 480 x 640 |
| 4      | Grayscale | 512 x 512 |
| 5      | Grayscale | 360 x 500 |
| 6      | Grayscale | 1080 x 1920 |
| 7      | Color    | 512 x 512 |
| 8      | Color    | 2835 x 4472 |
| 9      | Color    | 459 x 736 |
| 10     | Color    | 1200 x 1920 |

After the image is selected, the next process is the reading of values from pixels one by one. The values of a pixel which consists of red, green and blue, are separated into three values that are created in each color frame values. To make more data, the function of dividing frames is not just for one value for one pixel.

There are three values that can be made for the message insertion media. Frame division is based on three basic colors in each pixel that is Red, Green, and Blue. Next, the decimal values of color are changed in to the character based on the ASCII code. After getting the values, then the next process is the calculation
of the number of characters that can be input into the image that is distinguished from the colors obtained. The color values of the character are adjusted based on color matching with the value of the character that has been converted into a decimal. The number of characters are calculated to fit an image to a message in Indonesian text. The steps to adjust the values of the image with the distribution of Indonesian letter are as follows:

1. Take the greatest number of characters in the image. The distribution of A in Indonesian language reached 19.81%.
2. Once the number of characters A is held, the next process is to count how many number of characters that can be inserted by dividing the number of characters A to the percentage of the letter A.
3. Adjust the other characters with the percentage of each individual character. It should be in accordance with the distribution table of the letter. There should be no character that in number is less than the number of minimum letters based on the percentage in the distribution table. If one of the characters is not suitable, then the length of the character is adapted to the number of letters of the characters that are not suitable.
4. If the number of letter b is only 5 and that is not enough to meet the number of characters that previously obtained, then the length of characters that can be inputted that is adjusted with the letter b namely $5 / 2.64 \times 100 = 189$ characters. Further adjustments to the other letters in the same way.
5. Once the total of the letter percentage number is corresponding to a character length, further adjustments to the number of words which, according to research conducted in [4], the average use of the letters in each word is 6.6 letters.

Some calculation is performed to find out the number of characters that can be used as the basis for calculating the amount of code numbers of each letter that must be in the image as follows:

$$\text{Amount of Characters} = \frac{\text{Numbers of Character A}}{19.81 \times 10}$$

If there is a smaller number of letters the percentage multiplied by the number of characters, the number of characters must be adapted to the number of letters in the letter.

$$\text{Amount of Characters} = \frac{\text{Amount of Letters}}{\text{Percentage of Letters} \times 100}$$

The calculation of the number of words is:

$$\text{Amount of Letters} = \frac{\text{Amount of Characters}}{6.6}$$

### 4. Result

Here are the results of testing on four samples of a color image using an image size of 1200 X 1920. From the data in table 1, the results of testing from samples of 4 color images shows that these images have the letter A with the total number of 51,989. From this data, the percentage of letters that must be met by the next letter is as described in table 2.

#### Table 2. Results of Sample 4 Image Color Adjustment

| Letter | Amount | % Letter Dis | Number of Character | % Number Of Letter |
|--------|--------|--------------|---------------------|--------------------|
| A/a    | 51989  | 19.81%       | 262437              | 100.00%            |
| B/b    | 50889  | 2.97%        | 262437              | 97.88%             |
| C/c    | 51345  | 0.67%        | 262437              | 98.76%             |
| D/d    | 51303  | 3.72%        | 262437              | 98.68%             |
| E/e    | 54282  | 8.18%        | 262437              | 104.41%            |
| F/f    | 53018  | 0.25%        | 262437              | 101.98%            |
| G/g    | 51544  | 3.96%        | 262437              | 99.14%             |
| H/h    | 52345  | 1.97%        | 262437              | 101.07%            |
| I/i    | 52000  | 8.08%        | 262437              | 100.02%            |
| J/j    | 53685  | 0.85%        | 262437              | 103.26%            |
| K/k    | 52892  | 5.35%        | 262437              | 101.74%            |
| L/l    | 52439  | 3.50%        | 262437              | 100.87%            |
| M/m    | 50332  | 4.85%        | 262437              | 96.81%             |
| N/n    | 49996  | 9.56%        | 262437              | 96.17%             |
In the letter A with a percentage of 19.81% of the total number of letters 51,989 it is obtained the number of characters as much as 42,336 characters. By adjusting to the next letter, the number of through the letters in the sample 2 of color images is greater than the number of characters from the letters with the greatest percentage. So the number of characters that can be accommodated by the image is 262,437 characters and also 39,762 words. The results of the implementation for all images are obtained as described in Table 3.

### Table 3. Entire Sample Test Result

| Number | Type    | Size   | Character | Letter |
|--------|---------|--------|-----------|--------|
| 1      | Monochrome | 256 x 256 | 0         | 0      |
| 2      | Monochrome | 450 x 250 | 0         | 0      |
| 3      | Monochrome | 480 x 640 | 0         | 0      |
| 4      | Grayscale | 512 x 512 | 0         | 0      |
| 5      | Grayscale | 360 x 500 | 24,426    | 3700   |
| 6      | Grayscale | 1080 x 1920 | 314.386  | 46.278 |
| 7      | Color    | 512 x 512 | 0         | 0      |
| 8      | Color    | 2835 x 4472 | 1.357.485 | 205.679 |
| 9      | Color    | 459 x 736  | 42.336    | 6.414  |
| 10     | Color    | 1200 x 1920 | 262.437  | 39.762 |

The analysis of the research results with respect to the absorptive performance of the images to the messages using Indonesian languages is an important thing to close the discussion. The accommodation and fit rating of the letters to the image use the decimal value of the primary color (Red, Green, Blue) of each image pixel and the decimal value of the letter (ASCII) [8]. Data distribution of letters and the use of letters in the word and the text using Indonesian language for the adjustment are data from previously conducted research with 2,042,616 characters and 306,359 words (6.6 letters per word). The monochrome images cannot accommodate the character because the images only have values of 255 and 0. The digital images with greater resolution and use of many colors will accommodate more words in Indonesian language.

5. Conclusion

Based on the discussion and testing on the research, it can be concluded as follows:

1. The number of letters obtained from matching the color decimal value with a decimal value of the ASCII letters must meet the percentage of the distribution table of letters for all types of letters. If there is one letter that the number is less, then the maximum number of letters to the Indonesian language text to be readjusted based on the percentage of letters less.

2. The binary image cannot accommodate an Indonesian-language text with matching colors.
3. The results of the study are the result of matching the color values of samples with the distribution of letters in the Indonesian language.

6. References

[1] Khalil C and Hilmat F 2011 Combining Steganography and Cryptography International Journal on New Computer Architectures and Their Application 1(1) pp 199-208
[2] Atawneh S, Almomani A and Sumari P 2014 Steganography in digital images: Common approaches and tools IETE Technical Review 30(4) pp 344-358
[3] Muhammad K, Ahmad J, Farman , Jan Z, Sajjad M and Baik S W 2015 A secure method for color image steganography using gray-level modification and multi-level encryption KSII Transactions on Internet and Information 9(5) pp 1938-1962
[4] Andana G 2010 Analisis Frekuensi Pada Teks Berbahasa Indonesia Dan Modifikasi Algoritma kriptografi Klasik. Makalah IF3058 Kriptografi Tahun 2010 pp 9-18
[5] Suhad M And Shareef 2017 A Novel Approach for Arabic Text Steganography Based on the “BloodGroup” Text Hiding Method Engineering Technology & Applied Science Research 7(2) pp 1482-1485
[6] Promadi 2012 Perbedaan Semantik Antara Bahasa Indonesia Dan Bahasa Malaysia: Satu Kajian Awal Upaya Mengelak Kesalahpahaman Dan Perbedaan Budaya Antara Bangsa Serumpun Di Asia Tenggara Jurnal Sosial Budaya 9(2) pp 261-282
[7] Hamzah A 2010 Deteksi Bahasa untuk dokumen teks berbahasa indonesia, Seminar Nasional Informatika 2010 UPN "Veteran" Yogyakarta p. A5-A13
[8] Zhang Q and Wei X 2013 RGB Color Image Encryption Method Based on Lorenz Chaotic System and DNA Computation IETE Technical Review 30(5) pp 404-409