2017 International Seminar
on Application for Technology of Information and Communication (iSemantic)

ISBN : 978-1-5386-3085-3
IEEE CATALOG NUMBER : CFP17CUE-ART

UNIVERSITAS DIAN NUSWANTORO
UDINUS
SEMARANG

Technical co-Sponsor :
IEEE

ISBN :
978-1-5386-3085-3
IEEE CATALOG NUMBER :
CFP17CUE-ART

iSemantic

2017 International Seminar on Application for Technology of Information and Communication (iSemantic)

PROCEEDINGS

Empowering Technology for a Better Human Life

Email:
isemantic@lppm.dinus.ac.id
Website:
http://isemantic.dinus.ac.id/

Semarang | October 7th - 8th, 2017

Organized by

UDINUS
UNIVERSITAS DIAN NUSWANTORO • SEMARANG •

LPPMUDINUS

Lembaga Penelitian dan Pengabdian Masyarakat Universitas Dian Nuswantoro Semarang

The Faculty of Engineering Universitas Dian Nuswantoro Semarang

PROCEEDINGS

2017 International Seminar on Application for Technology of Information and Communication (iSemantic)

Empowering Technology for a Better Human Life

October 7th – 8th, 2017
Universitas Dian Nuswantoro
Semarang, Indonesia

ISBN: 978-1-5386-3085-3
IEEE Catalog Number: CFP17CUE-ART

Implementation of K-NN Based on Histogram at Image Recognition for Pornography Detection

Safira Nuraisha
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
nuraishafirak @gmail.com

Fandy Indra Pratama
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
fandyindra0794@gmail.com

Avira Budianita
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
viraanitaa@gmail.com

M. Arief Soeleman
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
arief22208@gmail.com

Abstract *– The development of information technology today has a positive and negative impact. One of the negative impacts is the spread of images containing inappropriate content (porn) uncontrollably so that it can be accessed by users from all walks of life, especially minors. There are several techniques to control the negative impact of the development of information technology, and one of them is by utilizing digital image processing in recognizing and detecting an image containing pornographic content. The technique used in this research is color segmentation on the image with YCbCr color model and classified and look for the similarity of training data with data testing using K-NN algorithm. The results obtained in this study using European and Asian skin color samples show that the prototype model designed can be used to detect pornographic, semi-pornographic or non-pornographic images with 90% accuracy.*

Keywords: Skin Detection, Pornography, K-NN, YCbCr, Color Segmentation

I. INTRODUCTION

Currently, access to information has become a very easy thing to do. Anyone from any age can access anytime and anywhere. However, not all information provided by the Internet is positive information, there is also negative information, and one of them is pornography. The problem of pornography has become a very serious issue today. Underage children can also access to the page or view pornographic content through the internet.

There are several techniques for detecting pornographic content, one of the techniques used for image segmentation is YCbCr color model. The use of the YCbCr color model on segmentation by Basilio et al. [1] is better than the RGB color model. However, the result of segmentation using thresholding on YCbCr color model showed 89.1% of accuracy. Based on that problem, this research will apply histogram to extract the image and use K-NN method as the introduction of the pornographic content image based on YCbCr based color segmentation.

II. RELATED WORK

Basilio et al. [1] use a new way method to detect explicit content images using YCbCr colors; this method can detect skin color. The main purpose using this method is to apply forensic analysis or detect pornographic images that exist on storage devices such as hard disks, USB memory, and so on. The results obtained using the proposed method will then be compared with the Paraben's Porn Detection Stick software which is one of the most used commercial devices for detecting pornographic images. In this study, Basilio et al. use 1000 sets of images consisting of 550 natural images and 450 images of highly explicit content. The results obtained are the proposed method of identifying up to 88.8% of the explicit content images, and 5% of false positives while the Paraben's Porn Detection Stick software achieves 89.7% of effectiveness for the same set of images but with 6, 8% of false positives. The investigators conclude that effectiveness of the proposed method in specific image detection is better than Paraben's Porn Detection Stick software.

Premal & Vinsley Research [2] proposed the YCbCr color model in image processing to detect forest fires. The proposed method adopts the color model rules due to the lack of complexity and effectiveness. The proposed method not only separates the flame pixels but also separates the high-temperature pixel of the fire center by taking into account the fire image statistics parameter in the YCbCr color space such as the mean and standard deviation. This method describes four rules for separating the fire area. Two rules are used to divide the fire area, and two other rules are used for high-temperature segmentation of the central fire area. The results obtained in comparison with other methods in the literature show a greater level of fire detection.

Riva Aktivia research [3] on the iris recognition using K-NN based Histogram method can produce accuracy in the left eye of 86,7% and right eye equal to 88,9% and the combination of both eyes achieve accuracy equal to 91,1%. So the Histogram-based K-NN algorithm is capable of being used for iris recognition.

Research about skin color detection to find out whether the image is a pornographic image or not done by Wibowo [4]. According to his research using HSV model (Hue Saturation Value) is used for human skin color segmentation and used to classify the image, especially regarding pornographic image classification. From the results of his research is still found a deficiency that if the background with a color similar to skin color will have an impact on detection. Background with a similar skin tone tends to increase the number of skin color pixels. Image brightness also tends to increase the number of pixels as skin color pixels.

III. METHODOLOGY AND IMPLEMENTATION

Problem-solving in this research consists of several stages. Figure 1 illustrates the overall process of this research. The process begins with reading the original image, then normalizing the RGB image followed by the image conversion process into the YCbCr, color model. The next process is the skin color segmentation and the value 1 (white color) for pixels that have a color similar to the color of the skin and the value 0 (black color) to vice versa. The last process in this research is the introduction of the image by using K-NN to determine the image including porn image, semi-porn, or not porn. Here is a more detailed explanation of the process in this study.

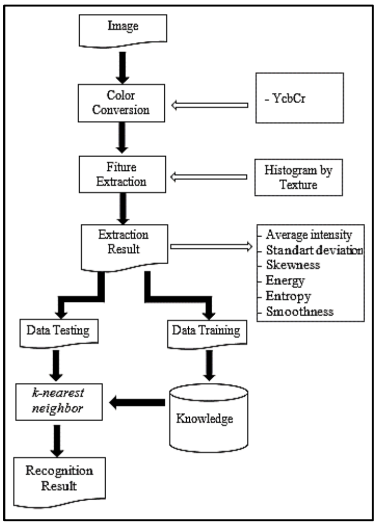

Fig 1. Flowchart

The flowchart referred to in Figure 1 is a groove of models intended to detect pornographic images. Here is an explanation of the model in this study.

3.1 Transforming RGB Color to YCbCr Color

In the first stage the original image is transformed from RGB color to YCbCr color, in which the RGB color has a certain color that has a pixel value range of 0-255, and if each color has a range from 0 to 255, the overall value of the image is 16,581,375 (16K) [5]. While YcbCr color has luma information represented by component Y which has 8-bit value between 16-235 and color information is stored with the component of Cb and Cr which is chroma difference between blue and red color having value between 16-240.

Here is the mathematical transformation from RGB color to YCbCr color.

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.481 & 128.553 & 24.966 \\ -37.797 & -74.203 & 112 \\ 112 & -93.786 & -18.214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

YCbCr produces a good performance to detect skin color compared to RGB format or the like [6].

3.2 Skin Thresholding

The process before the extraction of the image is the process of skin color segmentation is useful to get the skin of the object color has been transformed to color YCbCr. The thresholding formula for skin part identification is as follows [7]:

$$77 <= Cb \leq 127 \,\&\&\, 133 \leq Cr \leq 173$$

From this process, it will produce an image with skin threshold. Before the classification process, the color with the skin threshold is marked on each pixel. Pixels included in the range of skin threshold values are 0 (black), whereas for pixels not included in the skin threshold value range 255 (white). Clear pixel color differences (such as black and white) will make it easier to process the classification.

3.3 Histogram

In the study of a digital image, it is necessary to perform feature extraction useful to obtain a characteristic of the image, in obtaining the characteristics of an image there is a need to have a certain method used to get the value of each of these characteristics. In this study to get a characteristic value of the image using histogram texture extraction, the step used to obtain the image characteristics based on the histogram color; image histogram provides information that describes the spread of the intensity of the values on each part of the image.

The characteristic extraction method in each digital image has its extraction features; the histogram also has extraction features of the image characteristic histogram texture: skewness, entropy, energy, smoothness, standard deviation, and average intensity. Here is the formula for finding characteristics based on the histogram texture:

Average intensity (m) formula :

$$m = \sum_{i=0}^{L-1} i.p(i) \qquad (1)$$

where,
m = average intentity
i = gray level
p(i) = probability *i*
L = gray level max value

Standart deviation formula :

$$\sigma = \sqrt{\sum_{i=0}^{L-1} (i-m)^2 \, p(i)} \qquad (2)$$

Where,
σ = standart deviation
p(i) = opportunity function
i = gray level
m = average intensity

Skewness formula :

$$S = \sum_{i=0}^{L-1} (i-m)^3 \, p(i) \qquad (3)$$

Where,
p(i) = opportunity function
i = gray level
m = average intensity
L = grey level max value

Energy formula :

$$E = \sum_{i=0}^{L-1} [p(i)]^2 \qquad (4)$$

Where,
p(i) = opportunity function
i = gray level
L = grey level max value

Entropy formula :

$$e = \sum_{i=0}^{L-1} p(i) \log_2\big(p(i)\big) \qquad (5)$$

Where,
p(i) = opportunity function
i = gray level
L = grey level max value

Based on the value of skewness, entropy, energy, smoothness, standard deviation, and average intensity will be used as a characteristic of the image. So in the process of finding the similarity of data used to recognize images of pornography or not is based on the similarity of skewness, entropy, energy, smoothness, standard deviation, and average intensity of testing data and training data.

4.5 K-Nearest Neighbor

In the process of looking for similarities between testing data and training data using K-Nearest Neighbor algorithm. K-Nearest Neighbor (k-NN or KNN) algorithm is a method for classifying objects based on learning data closest to the object. K-Nearest Neighbor based on the concept of "learning by analogy". Data learning is described with n-dimensional numeric attributes. Each data learning represents a point, denoted by c, in the n-dimensional space. If a data query whose label is not known is inputted, K-Nearest Neighbor will search for the k-data of learning distance closest to the query data in the n-dimensional space. The distance between the query data and the learning data is calculated by measuring the distance between the points representing the query data with all points representing the learning data with the Euclidean Distance formula [8].

$$d_{12} = \sqrt{\sum_{k=1}^{n} (dx-dy)^2} \qquad (6)$$

The algorithm for computing the K-nearest neighbors is as follows [9]:

1. First, determine the parameter K = number of nearest neighbors.
2. Calculate the distance between the query-instance and all the training samples. For distance estimation, Euclidean distance method is used.
3. Sort the distance for all the training samples and determine the nearest neighbor based on the K-th minimum distance.
4. Get all the categories of the training data for the sorted value which falls under K.
5. Use the majority of nearest neighbors as the prediction value.

4.6 Accuracy Calculation

In search of accuracy in a test in the amount of data testing more than 1 data using the formula as follows:

$$\text{Accuracy} = \frac{Correct\ Number}{Total\ Data} X\ 100 \qquad (7)$$

From the formula will generate accuracy value to the data that has been done testing.

IV. EXPERIMENTAL RESULT

This study uses public dataset obtained from the internet (Google Image), with a total of 60 images consisting of 2 images containing no porn content, two semi-pornographic images and two images containing porn content used for training data and 54 remaining images used for data Testing.

Based on a questionnaire which is a category of pornography is a human image that does not wear clothes and visible body parts (breast and vagina). The semi-pornographic category of humans who wear minimal clothing, such as the human picture wearing bikini clothes. While that is not pornography is a human image wearing clothes from shoulder to knee, like a picture of a human who wore office clothing.

TABLE 1. Raw Data Sample

	Pornography
	Semi Pornography
	No Pornography

5.1 Segmentation

After the transformation from RGB color to YCbCr color, here is a sample of skin color segmentation calculation: The first known pixel has the YCbCr value as follows:

Y=8 Cb=100 Cr=153

Based on the value applied skin color segmentation formula with the threshold as follows:

$$77 <= Cb \le 127 \&\& 133 \le Cr \le 173$$

Then the result of that value is the skin category because the Cb value is between 77 and 127 and the Cr value is between 133 and 173.

After obtaining the result of the color segmentation process the labeling is done by giving a value of 0 on the pixel which is part of the skin and 255 on the part which is not considered as the skin.

Gambar 1. Segmentation Colour Sample

5.2 Extraction Using Histogram

With the results of segmentation is done extraction feature in order to find the characteristics of the image. The extraction features in this study are histogram-based to find the value of skewness, entropy, energy, smoothness, standard deviation, and the mean intensity which values can be used as characteristics of the image.

Here is the value of skewness, entropy, energy, smoothness, standard deviation, and the mean intensity generated from figure 3.

Table 2. Result Extraction Sample

Mu	Deviation	Skewness	Energy	entropy	smoothness
0,85535	0,35174	-1,3522	0,75255	0,41331	1,90272

5.3 K-NN

Furthermore, the calculation of the similarity of test data with train data using K-NN algorithm with K = 1 to obtain pornographic content information, semi pornography or not pornography.

Here are the extracts from the trained data obtained by taking 6 images of the total data with 2 pornographic content, 2 semi-pornographic images, and 2 non-pornographic images:

Table 3. Pornography Category

Mu	Deviation	Skewness	Energy	entropy	smoothness
0,85535	0,35174	-1,3522	0,75255	0,41331	1,90272
0,80356	0,39730	-1,4738	0,68429	0,49542	2,42752

Table 4. Semi-Pornography Category

Mu	Deviation	Skewness	Energy	entropy	smoothness
0,76259	0,42549	-1,4622	0,63791	0,54807	2,78421
0,67289	0,46915	-1,1705	0,55978	0,63210	3,38493

Table 5. Not Pornography Category

Mu	Deviation	Skewness	Energy	entropy	smoothness
0,93757	0,24193	-7,8774	0,88293	0,23359	9,00126
0,91596	0,27744	-9,8483	0,84604	0,28852	1,18380

The training data is used to reference the identification of pornographic images by looking for the similarity of test data with available trainer data. Example test data:

Table 6. Data Testing Sample

Mu	Deviation	Skewness	Energy	entropy	smoothness
0,80025	0,39980	-1,4761	0,68030	0,50005	2,45824

From the test data is done search similarity between test data and train data by using the calculation of K-NN.

From Table 6, the value of Mu, deviation, skewness, energy, entropy, and smoothness obtained from data testing is done nearest distance calculation using K-NN algorithm. With value, K = 1 with the result obtained that is 0,00739 is the minimum value of data testing in pornography category, 0.07862 minimum value of semi-pornography category and 0.31713 minimum value of non-pornographic category. Based on these values can be taken the smallest value as the most accurate value. So that the data testing is a category image pornography.

Based on the experiments the overall model that we propose this yields an accuracy rate of 90%.

V. CONCLUSION AND FUTURE WORK

Image detection containing pornographic content has been widely studied. In this study, researchers used skin color segmentation process, feature-based histogram extraction and image recognition using K-NN. The image recognition process shows better results using histogram-based feature extraction and the K-NN algorithm as an introduction. The accuracy of this proposed model yields an accuracy of 90%.

In this research, the result of accuracy is high enough for the detection of pornography, it is hoped for further research can improve the accuracy result in image detection. This research is limited only in Asian and European skin detection, for further research is expected to detect all skin types from different parts of the world. It is also hoped that further research can use the development of this research to detect other images (other than skin detection)

References

[1] J. A. M. BASILIO, G. A. TORRES, G. S. PÉREZ, L. K. T. MEDINA dan H. M. P. MEANA, “Explicit Image Detection using YCbCr Space Color Model as Skin Detection,” *Applications of Mathematics and Computer Engineering,* pp. 123-128.

[2] C. E. Prema dan S. S. Vinsley, “Image Processing Based Forest Fire Detection using YCbCr Colour Model,” *International Conference on Circuit, Power and Computing Technologies [ICCPCT],* pp. 1229-1237, 2014.

[3] R. Aktiva, “PENGENALAN IRIS MATA MENGGUNAKAN ALGORITME K-NEAREST NEIGHBOR BERBASIS HISTOGRAM,” IPB, Bogor, 2012.

[4] J. S. Wibowo, “Deteksi dan Klasifikasi Citra Berdasarkan Warna Kulit Menggunakan HSV,” *Jurnal Teknologi Informasi DINAMIK,* vol. 16, no. 2, pp. 118-123, 2011.

[5] E. R. Ariyanto, Wijanarto dan Sudaryanto, “KLASIFIKASI CITRA PORNO DENGAN ALGORITMA C4.5 BERBASIS MODEL WARNA YCbCr DAN SHAPE DETECTOR,” *Techno.COM,* vol. 15, no. 2, pp. 92-98, 2016.

[6] N. Trisnadik, A. Hidayatno dan R. R. Isnanto, “PENDETEKSIAN POSISI PLAT NOMOR KENDARAAN MENGGUNAKAN METODE MORFOLOGI MATEMATIKA”.

[7] R. Kusumanto, S. W. Pambudi dan A. N. Tompunu, “Aplikasi Sensor Vision Untuk Deteksi Multiface dan Menghitung Jumlah Orang,” *Semantik,* 2012.

[8] G. Kukharev dan A. Novosielski, “Visitor identification elaborating real time face recognition system,” *Winter School on Computer Graphics (WSCG),* p. 157 – 164, 2004.

[9] D. R. Muralidharan, “Object Recognition Using K-Nearest Neighbor Supported By Eigen Value Generated From the Features of an Image,” *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 2, no. 8, pp. 5521-5528, 2014.

[10] S. MI, "Pengenalan Wajah Menggunakan K-Nearest Neigbor dengan Praproses Transformasi Wavalet," *Jurnal Paradigma,* pp. 159-172, 2009.

[11] "Google Image," [Online]. Available: https://images.google.com.