

KLASIFIKASI CITRA PORNO DENGAN ALGORITMA C 4.5 BERBASIS MODEL WARNA YCbCr DAN SHAPE DETECTOR

Erwin Rizki Ariyanto¹, Wijanarto², Sudaryanto³

^{1,2}Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro

³Sistem Informasi, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro

Jl. Nakula I No. 5-11, Semarang, 5013, 024-3517261

E-mail : erwinrizki24@gmail.com¹, wijanarto@dsn.dinus.ac.id², sudaryanto@dsn.dinus.ac.id³

Abstrak

Konten internet dapat berupa pornografi atau jebakan pornografi dengan tujuan korban anak di bawah umur. Tahun 2010 sebesar 97% remaja pernah mengakses content pornografi, dimana hal tersebut berpengaruh negatif terhadap perkembangan anak dan remaja. Dengan demikian akses terhadap pornografi oleh anak perlu dijaga. Beberapa penelitian telah menghasilkan aplikasi pendeteksi pornografi dengan berbagai metode, seperti model warna YCbCr, shape descriptor yang saling mandiri. Dalam tulisan ini kami menyajikan klasifikasi citra porno dengan menggunakan algoritma C 4.5 dan Shape descriptor berbasis model warna YCbCr. Klasifikasi dengan C 4.5 merupakan tambahan metode perbaikan pada Shape Descriptor dan model warna YCbCr dan di harapkan dapat secara presisi mengklasifikasikan citra porno dan bukan. Hasil dari percobaan terhadap 40 citra 8 bit dengan dimensi 256X256 yang terbagi menjadi citra porno, berbikini, mug shots dan non-porno, metode di atas dapat mengklasifikasikan True Positives (TP) 16, False Positives (FP) 10, False Negatives (FN) 4, True Negatives (TN) 10 dengan akurasi 65%, error rate 35%, precision 0,615, recall 0,8 serta nilai Root Mean Squared Error (RMSE) 0,59. Semakin banyak data training semakin akurat hasil dari testingnya, walaupun dalam beberapa kasus metode di atas belum akurat dan hanya mengenail 3 dari 10 citra mug shots atau sebesar 43%.

Kata Kunci: c 4.5, shape descriptor, model warna ycbcr, citra pornografi

Abstract

Internet content can be in the form of pornography or pornographic trap with the aim of a minor victim. In 2010 amounted to 97% of adolescents ever accessing pornographic content, in which case the stretcher a negative effect on the development of children and adolescents. For those reasons, access to child pornography by the need to be on guard. Several studies have resulted in the application of pornography detection by various methods, such as YCbCr color models, the shape descriptor that is mutually independent. In this paper we present a pornographic image classification algorithms using C 4.5 and Shape descriptor-based YCbCr color model. Classification by C 4.5 is an additional method of repair on Shape Descriptor and YCbCr color model and is expected to precisely classify pornographic images instead. Results of the trial against the 40 images of 8 bits with dimensions of 256x256 which is divided into the image of porn, bikini, mug shots and non-porn, the above method can classify True Positives (TP) 16, False Positives (FP) 10, False Negatives (FN) 4, True Negatives (TN) 10 with an accuracy of 65%, 35% error rate, 0.615 precision, recall 0.8 and the value of the Root Mean Squared Error (RMSE) 0,59. The more training data more accurate results than testingnya, although in some cases the above method is not accurate and only about 3 out of 10 images mug shots or by 43%.

Keywords: c4.5, a shapedescriptor, YCbCrcolor models, the image ofpornography

1. PENDAHULUAN

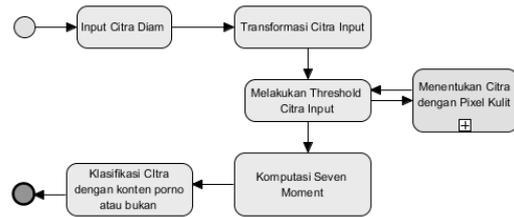
Internet berpengaruh secara signifikan pada perkembangan anak dan remaja, dalam hal akademik dan motivasi belajar, secara positif maupun negatif. Hampir 97% remaja pernah menonton atau mengakses materi pornografi [1]. Pornografi di internet bisa meliputi konten dewasa sampai dengan jebakan pornografi yang ditujukan kepada anak-anak di bawah umur [2].

Dalam menjaga akses internet terhadap pornografi baik secara umum maupun khusus pada anak dan remaja, beberapa penelitian terhadap citra pornografi telah dilakukan, misalnya menggunakan pendekatan RSOR dengan model warna HSV [3], lokalisasi kulit dengan ROIs [4], deteksi kulit dengan YCbCr [5]. Model warna YCbCr juga di terapkan dalam deteksi pornografi yang berbasis pada *skin detector* dan pengenalan pola [6][7]. Sementara *shape descriptor* di pakai untuk ekstraksi kulit dan klasifikasi C 4.5 yang di pakai dalam [8], menunjukkan hasil yang bagus.

Dalam tulisan ini , kami akan mengajukan suatu metode yang dapat mengklasifikasian citra porno dengan algoritma C 4.5 dan *shape descriptor* berbasis model warna YCbCr, teknik diatas diharapkan dapat mendeteksi citra porno dengan tujuan untuk dapat dipakai dalam suatu aplikasi internet saat mengunggah file atau filter file terhadap citra porno.

2. METODE

Beberapa metode dan prosedur penyelesaian masalah yang akan di pakai dalam penelitian ini terbagi dalam beberpa tahapan. Prosedur penyelesaian yang dilakukan adalah sebagai berikut:

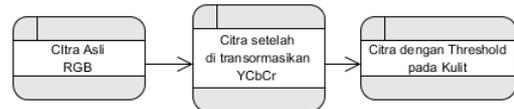


Gambar 1. Prosedur penyelesaian masalah

Dari gambar 1 tampak prosedur penyelesaian masalah di atas dapat di bagi menjadi 3 sekuen pertama, metode yang terdiri dari pendeteksian citra kulit, kedua, perhitungan *sevent moment* dan ketiga adalah mengklasifikasi citra, berikut detail metode yang diajukan untuk penyelesaian masalah :

2.1 Deteksi Kulit

Untuk mendeteksi kulit pada citra secara umum metode yang paling mudah dan cepat adalah dengan teknik transformasi warna dari warna citra input RGB menjadi YCbCr, seperti gambar model di bawah ini :



Gambar 2. Proses Skin Detection

Citra asli merupakan citra dengan model warna RGB, proses pada tahap ini citra akan ditransformasikan menjadi model warna YCbCr. Dalam [9] persamaan transformasi model warna RGB ke model warna YCbCr dengan menggunakan persamaan (1) seperti di bawah ini:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65,481 & 128,553 & 24,966 \\ -37,797 & -74,203 & 112,000 \\ 112,000 & -93,786 & -18,214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

Selanjutnya setelah melakukan transformasi adalah mengerjakan segmentasi yaitu dengan melakukan *skin thresholding* pada obyek orang dalam gambar tanpa memperhatikan warna kulit sehingga citra asli akan berubah menjadi citra dengan *skin thresold*. Nilai *range* untuk *skinthreshold* yang optimal seperti di

tunjukkan pada persamaan (2) dalam [5] adalah:

$$77 \leq Cb \leq 127 \text{ dan } 133 \leq Cr \leq 173 \quad (2)$$

Nilai Cg tidak dipakai karena tidak secara signifikan berpengaruh pada warna kulit seseorang pada citra diam.

2.2 Shape Descriptor

Metode *shape descriptor*, merupakan metode yang dipakai untuk menentukan obyek yang diinginkan dalam suatu citra. Hasil citra dari proses *skin detection* akan diproses kembali dalam metode *shape descriptor*. *Shape* merupakan karakter penting pada suatu obyek. Tujuan dari *shape descriptor* ini adalah untuk mengambil obyek yang ada pada suatu citra [8].

Shape descriptor yang digunakan adalah *seven normal moment invariants*. Langkah pertama dalam *shape descriptor* menurut [10] adalah menentukan nilai *moment of order (p+q)* yang didefinisikan sebagai berikut:

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \quad (3)$$

Nilai m_{pq} merupakan hasil dari *moment order (p+q)* dimana $p, q = 1, 2, 3, \dots$

Dalam [10] *central moment* didefinisikan sebagai berikut:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \quad (4)$$

dimana $\bar{x} = \frac{m_{10}}{m_{00}}$ dan $\bar{y} = \frac{m_{01}}{m_{00}}$

Normalized central moment dinotasikan sebagai η_{pq} , didefinisikan sebagai berikut:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma} \quad (5)$$

dimana $\gamma = \frac{p+q}{2} + 1$ dan $p+q = 2, 3, \dots$

Langkah berikutnya, *seven moment invariants* dapat dihitung dengan persamaan berikut :

$$\begin{aligned} \phi_1 &= \eta_{20} + \eta_{02} \\ \phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\ \phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\ \phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\ \phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (6) \\ \phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ &\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\ \phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &\quad + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned}$$

2.3 Klasifikasi C4.5

Setelah melewati tahap *shape descriptor*, hasil perhitungan *seven moment invariant* digunakan untuk proses klasifikasi apakah citra tersebut mengandung konten pornografi atau tidak. Algoritma yang digunakan untuk klasifikasi adalah algoritma C4.5. Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut : pilih atribut sebagai akar, buat cabang untuk masing – masing nilai, bagi kasus dalam cabang, dan ulangi proses untuk masing – masing cabang sampai semua kasus pada cabang memiliki kelas yang sama [11,13].

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut – atribut yang ada. Untuk menghitung *gain* digunakan rumus berikut :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i) \quad (7)$$

S sebagai himpunan kasus, A merupakan atribut, n merupakan jumlah partisi atribut A, $|S_i|$ adalah jumlah kasus pada partisi ke i, dan $|S|$ adalah jumlah kasus dalam S.

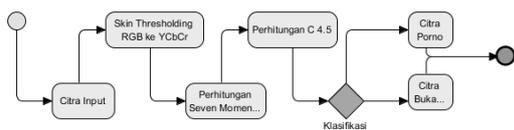
Sedangkan untuk menghitung *entropy* dapat dilihat pada rumus berikut dalam [11,12] :

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p_i \quad (8)$$

S merupakan himpunan kasus, n adalah jumlah partisi S, dan p_i adalah proporsi dari S_i terhadap S dengan S_i merupakan jumlah kasus pada partisi ke i.

Setelah semua perhitungan selesai maka

didapatkan suatu pohon keputusan yang bisa digunakan untuk mengklasifikasikan citra apakah mengandung konten pornografi atau tidak, berdasarkan *rules* yang dibentuk oleh pohon keputusan. Berikut gambar 3 merupakan *overall* proses dari klasifikasi untuk menentukan apakah citra mengandung konten porno atau bukan.



Gambar 3: Overall proses

3. HASIL DAN PEMBAHASAN

3.1 Analisis Data Training

Dalam rangka melakukan pelatihan (training) diperlukan suatu data training yang akan digunakan dalam algoritma klasifikasi C4.5 untuk mengklasifikasikan suatu citra apakah termasuk dalam kategori porno atau bukan porno.

Data training yang digunakan dalam pelatihan ini berjumlah 300 data citra diam, yang dibagi menjadi 4 kategori yaitu 75 citra porno, 75 citra bikini, 75 citra *mug shots*, dan 75 citra bukan porno. Perlakuan pertama dari data citra adalah mentransformasikan citra RGB menjadi YCbCr dan setelah itu melakukan perhitungan dengan *seven moment invariant*.

Langkah berikutnya, setelah melakukan tahap transformasi dan perhitungan *seven moment invariant*, adalah mengklasifikasikan data citra tadi dengan menggunakan algoritma C4.5. Dimana atribut yang digunakan dalam klasifikasi adalah hasil perhitungan *seven moment invariant*, dengan label yang dipakai sebagai hipotesis pornografi yang merupakan dugaan sementara penelitian untuk mengkategorikan suatu citra apakah

termasuk kategori porno atau bukan porno.

Hasil output yang dari algoritma C4.5 adalah suatu klasifikasi dalam bentuk pohon keputusan. Pohon keputusan ini memperlihatkan aturan dalam mengklasifikasikan apakah suatu citra mengandung konten pornografi atau tidak. Berikut tabel 1 yang merupakan rincian dari pohon keputusan yang terbentuk setelah melakukan klasifikasi data training :

Tabel 1: Tabel Hasil Pohon Keputusan

No	Variabel	Nilai
1.	Akar	Moment 4
2.	Jumlah Node	61
3.	Jumlah Daun	31
4.	Level	7

3.2 Analisis Data Testing

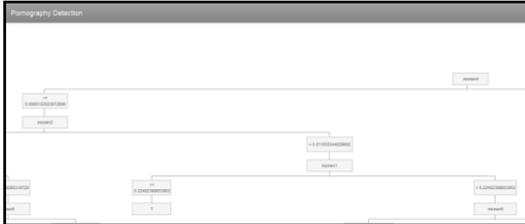
Dalam analisa terhadap data testing yang merupakan suatu proses untuk melakukan validasi terhadap data training, penelitian ini menggunakan data penelitian yang berjumlah 40 citra yang dibagi menjadi 4 kategori seperti dalam data training. Testing dalam algoritma klasifikasi C4.5 berdasarkan pada pohon keputusan yang terbentuk dari data training. Atribut dan label yang digunakan dalam data testing sama dengan atribut dan label yang digunakan dalam data training. Dalam tahap testing ini, suatu citra harus melalui tahap transformasi dan perhitungan *seven moment invariant* terlebih dahulu.

Setelah melalui tahap klasifikasi terhadap pohon keputusan yang terbentuk dari data training, berikut tabel 2 merupakan hasil dari klasifikasi data testing:

Tabel 2: Tabel Hasil Klasifikasi Data Testing

No	Kategori	Total	Σ Klasifikasi benar	Σ Klasifikasi salah
1.	Bikini	10	8	2
2.	Porno	10	8	2
3.	Mug shots	10	3	7
4.	Non porno	10	7	3

Tabel diatas menunjukkan jumlah klasifikasi benar dan jumlah klasifikasi salah untuk masing – masing kategori yang diuji. Klasifikasi benar berarti hasil klasifikasi sama dengan hipotesis, sedangkan klasifikasi salah menunjukkan bahwa hasil klasifikasi tidak sama dengan hipotesis.



Gambar 4. Halaman hasil klasifikasi

Gambar 4, merupakan *decision tree* hasil dari klasifikasi data training menggunakan algoritma C4.5 dan hasil pengujian metode yang mencakup *accuracy* dan *error rate*.

Gambar 5. Halaman hasil pengujian

Sementara gambar 5 di atas menampilkan *decision tree* hasil klasifikasi C4.5 dan hasil pengujian terhadap data testing.

3.3 Pengujian Data Training

Pengujian terhadap metode yang di pakai adalah dengan menggunakan tabel *confusion matrix* untuk menghitung akurasi, *error rate*, *precision*, dan *recall*. Dari data training diperoleh tabel *confusion matrix* berikut:

Tabel 3: Tabel Hasil Pengujian *Confusion Matrix Data Training*

No.	Variabel	Jumlah
1.	True Positives (TP)	140
2.	False Positives (FP)	55
3.	False Negatives (FN)	10
4.	True Negatives (TN)	95

Dari hasil tabel 3 di atas, maka *confusion matrix* tersebut diperoleh nilai akurasi, *error rate*, *precision*, dan *recall* sebagai berikut:

$$\begin{aligned}
 \text{akurasi} &= \frac{140 + 95}{300} * 100\% = 78,33\% \\
 \text{error rate} &= \frac{55 + 10}{300} * 100\% = 21,67\% \\
 \text{precision} &= \frac{140}{140 + 55} = 0,718 \\
 \text{recall} &= \frac{140}{140 + 10} = 0,93
 \end{aligned}$$

3.4 Pengujian Data Testing

Pengujian ini juga menggunakan tabel *confusion matrix* untuk menentukan akurasi, *error rate*, *precision*, dan *recall*. Dalam pengujian ini juga akan dihitung nilai *Root Mean Squared Error* (RMSE)[14]. Dari 40 data testing diperoleh hasil *confusion matrix* berikut:

Tabel 4: Tabel Hasil Pengujian *Confusion Matrix Data Testing*

No.	Variabel	Jumlah
1.	True Positives (TP)	16
2.	False Positives (FP)	10
3.	False Negatives (FN)	4
4.	True Negatives (TN)	10

Dari hasil *confusion matrix* tersebut, diperoleh nilai akurasi, *error rate*, *precision*, dan *recall* sebagai berikut:

$$\begin{aligned}
 \text{akurasi} &= \frac{16 + 10}{40} * 100\% = 65\% \\
 \text{errorrates} &= \frac{10 + 4}{40} * 100\% = 35\% \\
 \text{precision} &= \frac{16}{16 + 10} = 0,615 \\
 \text{recall} &= \frac{16}{16 + 4} = 0,8
 \end{aligned}$$

Untuk mendapatkan nilai RMSE, label diinisialisasi sebagai 1 dan 0. Angka 1 menunjukkan kasus “Ya”, sedangkan angka 0 menunjukkan kasus “Tidak”. Nilai RMSE yang didapatkan adalah sebagai berikut:

$$RMSE = \sqrt{\frac{14}{40}} = \sqrt{0,35} = 0,59$$

Hasil akurasi dan RMSE yang diperoleh diatas menggunakan data training yang berjumlah 300 gambar. Pengujian juga dilakukan menggunakan 100 data training dan 40 data testing untuk mengetahui apakah ada perbedaan akurasi terhadap banyaknya data training yang digunakan. Berikut hasil yang diperoleh dari pengujian tersebut:

Tabel 5: Tabel Perbandingan Akurasi

No	Jumlah Data Training	Jumlah Data Testing	Akurasi
1.	100	40	60%
2	300	40	65%

4. KESIMPULAN DAN SARAN

Dengan demikian berdasarkan analisis dan pengujian terhadap data yang di peroleh, maka dalam penelitian ini dapat disimpulkan sementara sebagai berikut:

1. Bahwa hasil penelitian menunjukkan metode yang digunakan dalam penelitian kurang bisa mengenali citra *mug shots*, dimana didapatkan jumlah klasifikasi benar 3 serta jumlah klasifikasi salah 7 yang berasal dari 10 data citra *mug shots* yang dipakai sebagai data testing.
2. Berdasarkan data testing sebanyak 40 terbagi dalam 4 jenis yaitu citra porno, citra wanita memakai bikini, citra *mug shots*, dan citra bukan porno diperoleh hasil komputasi sebagai berikut, *True Positives* (TP) sebanyak 16 citra, *False Positives* (FP) sebanyak 10 citra, *False Negatives* (FN) sebanyak 4 citra, *True Negatives* (TN) sebanyak 10 citra, dengan akurasi sebaesar 65%, *error rate* dengan nilai 35%, *precision* sebesar 0,615, *recall* dengan nilai 0,8 serta *Root*

Mean Squared Error (RMSE) bernilai 0,59.

3. Ditemukan kecenderungan dalam penelitian ini bahwa semakin besar data yang digunakan dalam pelatihan berpengaruh secara signifikan terhadap keakuratan metode. Juga semakin besar data yang dipakai dalam training, maka metode yang dipakai akan cenderung memiliki akurasi yang tinggi dalam mendeteksi citra pornografi. Training terhadap 100 data mendapatkan akurasi 60% sedangkan training terhadap 300 data mendapatkan akurasi 65% dari 40 data yang digunakan untuk testing.

DAFTAR PUSTAKA

- [1] Tri Suyatno, "Pengaruh Pornografi Terhadap Perilaku Belajar Siswa," *Jurnal Pendidikan Dompot Dhuafa*, vol. I, 2011.
- [2] Ridwan Sanjaya, *Parenting untuk Pornografi di Internet*. Jakarta, Indonesia: Elex Media Komputindo, 2010.
- [3] Pedro Ivan Tello Flores, Luis Enrique Colmenares Guillen, and Omar Ariosto Nino Prieto, "Approach of RSOR Algorithm Using HSV Color Model for Nude Detection in Digital Images," *Journal Computer and Information Science*, vol. 4, pp. 29-45, July 2011.
- [4] Sotiris Karavarsamis, Nikos Ntarmos, Kostantinos Blekas, and Ioannis Pitas, "Detecting Pornographic Images by Localizing Skin ROIs," *International Journal of Digital Crime and Forensics*, 2013.
- [5] Jorge Alberto Marcial Basilio, Gualberto Aguilar Torres, Gabriel Sanchez Perez, L. Karina Toscano Medina, and Hector M. Perez Meana, "Explicit Image Detection

- Using YCbCr Space Color Model as Skin Detection," 2011.
- [6] Bei-Bei Liu, Jing-yong Su, Zhe-ming Lu, and Zhen Li, "Pornographic Images Detection Based on CBIR and Skin Analysis," 2008.
- [7] Jiann-Shu Lee, Feng-Shuo Yu, and Kai-Yang Huang, "Pornography Detection Based on Morphological Features," 2013.
- [8] Qing-Fang Zheng, Wei Zeng, Gao Wen, and Wei-Qiang Wang, "Shape-based Adult Images Detection," 2004.
- [9] Charles Poynton, *Digital Video and HDTV Algorithms and Interfaces*. San Fransisco, United States of America: Morgan Kaufmann, 2003.
- [10] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing Second Edition*.: Prentice Hall, 2002.
- [11] Kusrini and Emha Taufiq Luthfi, *Algoritma Data Mining*. Yogyakarta: Andi Offset, 2009.
- [12] Daniel T. Larose, *Discovering Knowledge In Data*.: A John Wiley & Sons, 2005.
- [13] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining Concepts and Techniques*.: Morgan Kaufmann, 2012.
- [14] K. Pratheep Moses, "An Approach To Reduce Root Mean Square Error in Toposheets," *European Journal of Scientific Research*, vol. 91, pp. 268-274, November 2012.