Journal of Applied Informatics and Computing (JAIC)
Vol.2, No.2, Desember 2018, pp. 29~33
e-ISSN: 2548-6861

Analisis *Robustness* Teks Captcha Paypal HIP Menggunakan *Template Matching*

Fitrah Maharani Humaira [1]*, Tengku Musri [2], Sarimuddin [3]***,**
Dwi Samsuifin Alham [4]*, Aurista Miftahatul Ilmah [5]***
* Teknik Listrik Industri, Politeknik Negeri Madura
** Teknik Informatika, Politeknik Negeri Bengkalis
***Teknik Informatika, Universitas Sembilan Belas November Kolaka
humaira@poltera.ac.id [1], musri@polbeng.ac.id [2], sarimuddin85@gmail.com [3], aurista@poltera.ac.id [5]

Article Info

Article history:

Received : 08-11-2018
Revised : 27-11-2018
Accepted : 04-11-2018

Keyword:

CAPTCHA,
CAPTCHA paypall image processing,
robustness,
template matching

ABSTRACT

CAPTCHA refer to Completely Automated Public Turing test to tell Computers and Humans Apart. CAPTCHA are used to ensure that the operators are human not robots. The basic idea of using CAPTCHA is segmentation and recognition. Random characters, graphic images, or CAPTCHA audio become possible solutions to improve security and resilience for protection systems. In this paper used CAPTCHA random characters. However the CAPTCHA text needs to be analyzed again whether it is still solved by the computer or not it needs to be analyzed, improved, and developed to avoid automatic interference. Data set of text CAPTCHA paypal or so-called paypal HIP with 20 pieces of training data to get the template as much as 36 images that is from the numbers 0-9 and the letter A-Z. This particular paypal HIP data is limited by not using numbers 0 and 1 with the letters O and Q because of the similarity between the data. The method used starts from pre-processing, segmentation, and classification. Pre-processing techniques used consist of removing noise by tresholding and using cleaning techniques. We use bounding box and padding for segmentation method. And then for classification used counting pixel, vertical projections, horizontal projections, dan template correlation. By using these methods will be known which method can recognize CAPTCHA text accurately so as to affect the robustness of the CAPTCHA text.

I. PENDAHULUAN

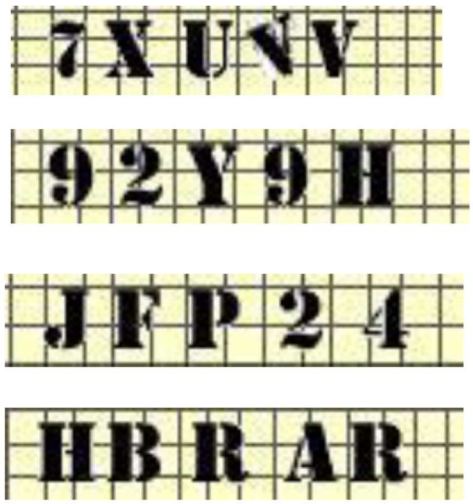
CAPTCHA merupakan kepanjangan dari Completely Automated Public Turing test to tell Computers and Humans Apart. CAPTCHA merupakan framework dari keamanan jaringan yang dapat membantu untuk menemukan cacat dan mencegah serangan dari internet [1]. CAPTCHA digunakan untuk memastikan bahwa yang mengoperasikan adalah manusia bukan robot. Ide dasar penggunaan CAPTCHA adalah segmentasi dan pengenalan (*recognition*). Karakter acak, gambar grafis, atau CAPTCHA audio yang menjadi solusi yang mungkin untuk meningkatkan ketahanan dan keamanan untuk sistem perlindungan [2]. Pada paper kali ini digunakan CAPTCHA karakter acak. Text CAPTCHA adalah skema CAPTCHA yang paling populer karena kemudahan konstruksi dan user friendly. Sejumlah besar skema teks singkat CAPTCHA yang diterapkan dapat dibongkar, seperti Google, Yahoo!, dan Microsoft. Desainer CAPTCHA biasanya belajar dari kegagalan sebelumnya untuk mendesain CAPTCHA dengan peningkatan keamanan dan kegunaan [4].

Lebih lanjut, tantangan CAPTCHA menjadi tidak aman karena dan mudah untuk menerobos teknik mutakhir CAPTCHA yang sudah ada [5]. Bagaimanapun teks CAPTCHA perlu diteliti lagi apakah masih dipecahkan oleh komputer atau tidak maka perlu dianalisis, diperbaiki, dan dikembangkan untuk menghindari gangguan otomatis. Dalam beberapa tahun terakhir, beberapa penelitian tentang pengenalan CAPTCHA dan aplikasi terkait disajikan satu demi satu untuk membahas keaamanan setiap jenis CAPTCHA. Sebagai contoh, Pada Penelitian terdahulu digunakan metode Recurrent Neural Network (RNN),

Support Vector Machine (SVM), dan Extreme Learning Machine (ELM) untuk pengenalan CAPTCHA [1]. Pada paper ini akan diteliti tentang kekekalan (*robustness*) dengan menggunakan template matching. Data yang digunakan adalah teks captcha paypal atau biasa disebut paypal HIP.

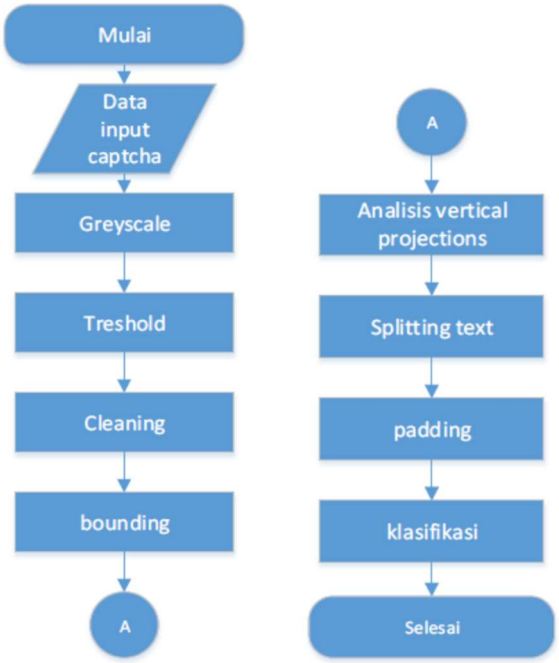
II. METODE PENELITIAN

Data set citra yang digunakan berupa teks CAPTCHA paypal atau biasa disebut paypal HIP dengan data training sebanyak 20 citra untuk mendapatkan template sebanyak 36 citra yaitu dari angka 0-9 dan huruf A-Z. Citra ini terdiri dari 5 karakter. Khusus data paypal HIP ini dibatasi dengan tidak menggunakan angka 0 dan 1 dengan huruf O dan Q karena kemiripan antara data tersebut. Data testing yang digunakan sebanyak 100 citra dengan contoh seperti pada Gambar 1.



Gambar 1. Contoh data testing

Metode yang digunakan dimulai dari pre prosesing, segmentasi, dan klasifikasi. Teknik pre prosesing yang digunakan terdiri dari menghilangkan noise dengan tresholding dan menggunakan teknik cleaning yang simpel. Metode segmentasi yang digunakan adalah bounding box dan padding. Terakhir metode klasifikasi yang digunakan adalah counting pixel, vertical projections, horizontal projections, dan template correlation. Diagram proses dapat dilihat pada Gambar 2.



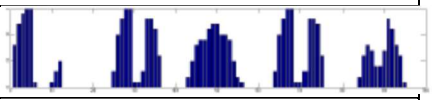
Gambar 2. Diagram proses analisis

Adapun tahapan proses pre prosesing dan segmentasi dapat dilihat pada Tabel 1. Tahapan dibagi menjadi 2 tahapan utama yaitu mulai dari tahap preprosesing, segmentasi. Hasil dari segmentasi ini yang akan masuk pada tahap klasifikasi. Pada setiap tahapan terdiri dari sub tahapan yang merupakan proses detail dari analisis kekekalan (robustness) teks CAPTCHA tersebut.

TABEL 1.

TAHAPAN PRE PROCESSING DAN SEGMENTASI TEKS CAPTCHA

Tahapan Utama	Sub Tahapan	Hasil Citra
Pre Processing	*Original Image*	
	Setelah di *grayscale*	
	Setelah di *treshold*	
	Hasil *cleaning*	
	Hasil *bounding*	
Segmentasi	Input ke *segmenter*	

Tahapan Utama	Sub Tahapan	Hasil Citra
	Analisis *Vertical Projections*	
	Memisahkan *image character*	
	Setelah *padding*	

Setelah mendapatkan image hasil segmentasi, tahap terakhir adalah klasifikasi. Pada tahap ini image hasil segmentasi diklasifikasi dengan berbagai metode. Metode yang digunakan pada paper ini adalah dengan menggunakan counting pixel, vertical projections, horizontal projections, dan template correlations. Adapun algoritma tiap metode klasifikasi adalah sebagai berikut [3]:

A. Counting pixel

PixelCount(I)

1. k ← 0
2. for r ← 1 to $I_{numRows}$
3. do for c ← 1 to I_{numCol}
4. do k ← k+I [r] [c]
5. Return k

ClassifyPC(I,T)

1. D ←Ø
2. For each Template $t_i \in T$
3. do $d_i \leftarrow abs(PixelCount(t_i) -$
4. $PixelCount(I))$
5. Return k such that $d_k = \min(D)$

B. Vertical projections

VerticalProjection(I)

1. V ← Ø
2. for r ← 1 to $I_{numRows}$
3. do for c ← 1 to I_{numCol}
4. do v_c ← v_c+I [r] [c]
5. Return V

Classify VP (I,T)

1. R ←Ø
2. For each Template $t_i \in T$
3. do $r_i \leftarrow Corr2($
4. $VerticalProjections(t_i)$,
5. $VerticalProjections(I)$,
6. Return k such that $r_k = \max(R)$

C. Horizontal projections

HorizontalProjection(I)

1. V ← Ø
2. for r ← 1 to $I_{numRows}$
3. do for c ← 1 to I_{numCol}
4. do v_r ← v_r+I [r] [c]
5. Return V

Classify HP (I,T)

1. R ←Ø
2. For each Template $t_i \in T$
3. do $r_i \leftarrow Corr2($
4. $HorizontalProjections(t_i)$,
5. $HorizontalProjections(I)$,
6. Return k such that $r_k = \max(R)$

D. Template correlations

1. R ←Ø
2. For each Template $t_i \in T$
3. do $r_i \leftarrow Corr2(t_i, I)$
4. Return k such that $r_k = \max(R)$

Pada paper ini akan dilakukan beberapa pengukuran yaitu confidence, akurasi karakter, akurasi HIP, dan waktu komputasi. Penjelasan mengenai setiap tahapan adalah sebagai berikut.

1. Confidence

Dari perhitungan confidence dilakukan perhitungan rata-rata confidence dari 100 data testing dan mencantumkan confidence minimum. Pengukuran confidence sendiri adalah sebagai berikut:

$$C = \prod_{i=1}^{5} r_i$$

2. Akurasi HIP

Akurasi HIP ini menghitung apakah pengenalan CAPTCHA paypal HIP dengan metode yang diusulkan ini benar atau tidak dilihat dari kelima karakter nya. Perhitungan dari akurasi HIP ini adalah sebagai berikut:

$$A_{HIP} = \frac{\sum HIP\ correct}{\#data\ testing} x100\%$$

3. Akurasi karakter

Akurasi karakter menghitung berapa banyak karakter benar yang telah dikenali dengan penggunaan berbagai metode klasifikasi yang telah diusulkan. Perhitungan dari akurasi karakter HIP adalah sebagai berikut:

$$A_{Char} = \frac{5x\ \#\ karakter\ yang\ benar}{5x\#\ data\ testing} x100\%$$

4. *Waktu komputasi*

Perhitungan ini dihitung berdasarkan waktu komputasi yang telah digunakan untuk mengenali citra CAPTCHA paypal HIP ini. Pengukurannya menggunakan second. Caranya menghitung waktu eksekusi dari 100 data testing.

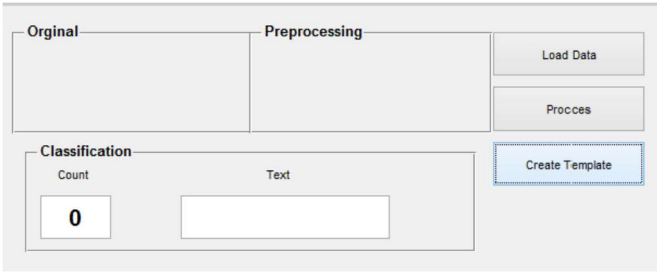
Informasi sistem komputer yang digunakan adalah dengan sistem operasi windows 8.1 pro 64-bit, Intel core i3 dengan memori 2GB dan matlab R2012a.

III. HASIL DAN PEMBAHASAN

Pada hasil dan pembahasan ini akan diuraikan hasil dari running dengan menggunakan matlab serta hasil perhitungan akurasi yang dengan keempat perhitungan yang telah dijelaskan pada bab metode. Hasil dari running matlab adalah sebagai berikut.

1. *CAPTCHA GUI*

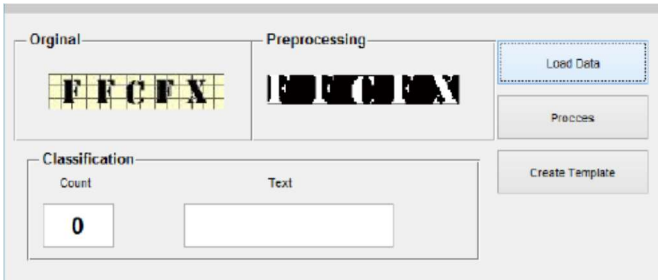
Gambar 3 merupakan *Graphical User Interface* (GUI) pada pengenalan CAPTCHA. GUI terdiri dari menu untuk menampilkan gambar original, kemudian tombol process digunakan untuk menampikan hasil dari preprocessing. Kemudian akan ditampilkan hasil klasifikasi dengan menghitung berapa karakter yang terbaca dan hasil dari pembacaan teks CAPTCHA.



Gambar 3. CAPTCHA GUI

2. *Load Data*

Pada saat memilih menu load data, dapat dipilih *image* mana yang akan dianalisis, kemudian akan muncul hasil dari preprocessing. Setelah dilakukan pre prosesing didapatkan data yang siap masuk ke tahap segmentasi. Gambar 4 menampilkan hasil dari load data.



Gambar 4. Hasil Load Data

3. *Hasil segmentasi*

Gambar 5 merupakan hasil dari segmentasi teks CAPTCHA yang akan dianalisis.

Gambar 5. Hasil Segmentasi

4. *Keseluruhan hasil*

Keseluruhan hasil dari system yang dibuat dapat dilihat pada Gambar 6. Pada GUI ditampilkan banyaknya huruf pada CAPTCHA yang terbaca, pada contoh yang digunakan yaitu CAPTCHA Paypal terdapat 5 karakter huruf. Hasil yang terbaca dapat dilihat pada menu "Text" pada "Classification".

Gambar 6. Hasil segmentasi dan klasifikasi

Dari hasil load data di atas didapatkan hasil pengenalan yang benar. Dari perhitungan *confidence* dan keakurasian didapatkan hasil seperti pada Tabel 2. Dari lima jenis akurasi yang digunakan, dapat disimpulkan bahwa klasifikasi menggunakan Horizontal Projections (HP) dan Template Correlations (TC) menghasilkan akurasi karakter dan akurasi HIP sebesar 100%. Untuk perhitungan confidence, metode klasifikasi yang paling menghasilkan nilai confidence paling besar adalah metode Horizontal Projections, sedangkan waktu paling cepat yang dibutuhkan untuk mengenali teks CAPTCHA adalah dengan metode Horizontal Projections.

TABEL 2.

HASIL PERHITUNGAN PENGUKURAN

Jenis Akurasi	PC	VP	HP	TC
Achar	64,8%	99,4%	100%	100%
AHIP	10%	97%	100%	100%
C avg	n/a	98,9%	98,8%	95,9%
C min	n/a	89,1%	93,9%	67,9%
T avg	35,18 s	20,16 s	19, 44 s	19,84 s

Keterangan:
PC = *Pixel Counting*
VP = *Vertical Projections*
HP = *Horizontal Projections*
TC = *Template Correlations*

IV. KESIMPULAN

Dari hasil perhitungan akurasi karakter dan akurasi HIP maka dapat disimpulkan bahwa metode klasifikasi yang paling akurat untuk mengenali teks CAPTCHA paypal HIP adalah metode horizontal projection dan template correlation dengan nilai akurasi sebesar 100%. Data set berupa teks CAPTCHA paypal atau biasa disebut paypal HIP dengan data training sebanyak 20 citra untuk mendapatkan template sebanyak 36 citra yaitu dari angka 0-9 dan hurup A-Z. Khusus data paypal HIP ini dibatasi dengan tidak menggunakan angka 0 dan 1 dengan huruf O, I dan Q karena kemiripan antara data tersebut. Data testing yang digunakan sebanyak 100 citra. Program yang digunakan adalah Matlab R2012a dengan tahapan-tahapan yang terdapat pada pengolahan citra yaitu pre-prosesing, segmentasi dan klasifikasi. Diperlukan metode lain untuk dapat menjawab pertanyaan apakah teks CAPTCHA masih bisa dipecahkan oleh komputer atau tidak sehingga perlu dianalisis, diperbaiki, dan dikembangkan untuk menghindari gangguan otomatis.

DAFTAR PUSTAKA

[1] Chen, C.J, You- Wei Wang, Wen-Pinn Fang, 2014, "A Study on CAPTCHA Recognition", Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing;

[2] Sakkatos, et. Al., 2014, "Analysis of Text-Based CAPTCHA Images using Template Matching Correlation Technique", The 4th Joint International Conference on Information and Communication Technology, Electronic and Electrical Engineering (JICTEE-2014)

[3] Kluever, K.A, 2008, "Breaking the Paypal HIP: A Comparison of Classifiers"

[4] Tang, M. et. Al., 2016, "Research on Deep Learning Techniques in Breaking Text-based CAPTCHAs and Designing Image-based CAPTCHA", IEEE Transactions On Information Forensics And Security, Vol. 14, No. 8

[5] Banday, M.T, and Shafiya A. S, 2014, "Service Framework for Dynamic Multilingual CAPTCHA Challenges: IN-CAPTCHA", International Conference on Advances in Electronics, Computers and Communications (ICAECC)