SIMULASI DAN ANALISIS PENGURAIAN GIMPY CAPTCHA MENGGUNAKAN PENGOLAHAN CITRA DAN SEGMENTASI HURUF

Yudha Ryandieka[1], Koredianto Usman[2], Suryo Adhi Wibowo[3]

[1]Teknik Telekomunikasi, Fakultas Teknik Elektro, Universitas Telkom

[1]yudharyandieka@gmail.com

Abstrak

CAPTCHA merupakan suatu program tes yang secara otomatis menguji dan membedakan antara manusia dan komputer dengan tujuan untuk mengatur akses ke website. CAPTCHA ditampilkan dalam program berbentuk tes Artificial Intelligence yang dapat berupa teks, gambar atau audio yang sulit untuk dikenali komputer tetapi dapat dikenali oleh manu sia. CAPTCHA berfungsi untuk mencegah spam yang berusaha memasuki sistem. Namun seiring perkembangan zaman, dikhawatirkan akan muncul sebuah program yang mampu memecahkan tantangan CAPTCHA. Oleh karena itu, penulis melakukan pemecahan tantangan CAPTCHA berbasis teks dengan jenis gimpy yang diimplementasikan di www.hotmail.com untuk mengecek seberapa kuat pertahanan CAPTCHA terhadap sebuah website.

Sistem ini bekerja dengan menampilkan CAPTCHA berjenis teks gimpy dengan karakteristik berupa huruf besar/kecil dan angka dengan distorsi berupa warping. Dalam proses penguraiannya, langkah awal meliputi preprocessing dengan melakukan konversi warna ke hitam putih, penghapusan piksel dengan ukuran luas yang kecil dan cropping. Dilanjutkan proses segmentasi karakter dengan memisahkan persebaran piksel yang saling terpisah satu sama lain, lalu proses normalisasi dengan melakukan rotasi, operasi morfologi dan resize ukuran karakter. Langkah terakhir, dilakukan proses ekstraksi ciri dengan metode square dan klasifikasi menggunakan metode Mean Square Error (MSE)

Dari simulasi ini dihasilkan akurasi terbaik untuk proses segmentasi sebesar 79,167% dengan tantangan CAPTCHA yang dipecahkan sebanyak 15 sampel dari total 120 sampel yang menghasilkan akurasi file CAPTCHA sebesar 12,5% dan akurasi karakter sebesar 70,74%. Waktu komputasi rata - rata untuk setiap proses pengenalan CAPTCHA berlangsung selama 0,156 detik.

Kata Kunci : CAPTCHA, Art ificial Intelligence, spam, preprocessing, websi te , cropping, resize, square, Mean Square Error

Abstract

CAPTCHA is a test program that automatically test and distinguish between human and computer with the goal to set up access to the website . CAPTCHA is shown in text, images or audio . They are difficult to be recognized by the computer but can be recognized by humans . CAPTCHA is used to prevent spam that tried to enter the system. But over the times, people worry if there is a program that can solve CAPTCHA challeng e . Therefore, the author conducted a text - based CAPTCHA solving challenges with Gimpy types that have been implemented in the website www.hotmail.com to check how robust CAPTCHA defenses against a website.

The security system works by dis pl aying Gimpy CAPTCHA text type which has in the form of upper / lower case and numbers with warping for distortion . In the breaking scheme , the initial step includes preprocessing pursued by conv erting color to black and white , eliminate small pixels and th en cropping . Followed by the character segmentation process to separate the pixels based on their distribution which mutually exclusive of each other, then the process of normalization include rotation , resize the character size and morphological operation s. The final step is feature extraction process using square method and classification using Mean Square Error (MSE).

From this simulation produced the best accuracy for segmentation is 79,167% and CAPTCHA cha llenges are solved as many as 15 samples from a total of 120 sample trials r esulting in an accuracy of 12,5 % CAPTCHA file with character accuracy of 70.64 % . The average computation time for each CAPTCHA recognition process lasts for 0.156 seconds.

Keywords : CAPTCHA , Art ificial Intelligence, spam , preprocessing, website , cropping, resize, square, Mean Square Error

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Di era komunikasi global seperti saat ini, kontrol dan keamanan di dalam jaringan internet merupakan suatu hal yang sangat penting. Oleh karena itu, dibutuhkan sesuatu yang mampu menghambat atau menghentikan gangguan-gangguan yang ada di dalamnya. Salah satu dari gangguan yang sudah populer dikenal oleh masyarakat luas adalah *spam*.

Blog pada umumnya menyediakan fasilitas interaksi antara *blogger* (pembuat *blog*) dan pengunjung melalui pengisian komentar. Karena populernya *blog*, media komentar ini sering dimanfaatkan untuk menaikkan popularitas *website-website* baru. Hal ini awalnya memang bukan masalah, tapi setelah itu *blog* mulai dipenuhi komentar-komentar yang tidak ada hubungannya dengan artikel yang ditulis dan juga iklan-iklan yang datang secara terus menerus tanpa henti.

Perilaku aneh ini adalah ulah robot *spam*. *Spam* merupakan penggunaan perangkat elektronik yang mengirimkan pesan secara bertubi-tubi tanpa dikehendaki oleh penerimanya. *Spam* dikirimkan oleh pihak-pihak yang memiliki tujuan untuk mengiklankan produknya dengan biaya operasional yang sangat rendah, karena *spam* tidak memerlukan daftar tertentu untuk mencapai para pelanggan-pelanggan yang diinginkan. Oleh karena mudahnya mengirimkan *spam* ke dalam suatu forum, maka banyak *spammers* yang muncul dengan jumlah yang sangat tinggi. Akibatnya, banyak pihak yang merasa tidak nyaman.

Untuk mengatasi hal tersebut, maka muncul teknologi baru yang dinamakan dengan CAPTCHA. Istilah ini merupakan akronim bahasa Inggris dari *Completely Automated Public Turing test to tell Computers and Humans Apart* yang merupakan program berbentuk tes yang secara otomatis menguji dan membedakan antara manusia dan komputer dengan tujuan untuk mengatur akses ke *website*.

Perbincangan mengenai tes terotomatisasi pertama kali dilakukan oleh Moni Naor dari *Weizmann Institute of Science* dan dituliskan dalam naskah berjudul *Verification of a human in the loop, or Identification via the Turing Test*. Nama CAPTCHA kemudian dipublikasikan pada tahun 1997 oleh Andrei Broder dan rekan-rekannya untuk mencegah *bot* menambahkan URL ke mesin pencari mereka. Mereka memutuskan untuk menambahkan tes ke halaman registrasi. Hingga pada tahun 2000, hak cipta CAPTCHA dipegang oleh tim profesor dari Universitas Carnegie Mellon, yakni : Luis von Ahn, Manuel Blum, dan Nicholas J. Hopper yang sebelumnya telah melakukan pengembangan pada CAPTCHA.

Jenis CAPTCHA yang seringkali digunakan pada beberapa *website* adalah CAPTCHA berbasis teks. CAPTCHA berbasis teks pun terbagi lagi dalam beberapa sub-jenis CAPTCHA dengan tingkat keamanan yang berbeda satu sama lain. Salah satu CAPTCHA berbasis teks yang populer digunakan di beberapa *website* adalah Gimpy CAPTCHA.

Penelitian terkait Gimpy CAPTCHA dilakukan oleh Shih-Yu Huang, Yeuan-Kuen Lee, Graeme Bell dan Zhan-he Ou yang didokumentasikan pada jurnal berjudul "*A Projection-based Segmentation Algorithm for Breaking MSN and Yahoo CAPTCHAs*". Algoritma yang digunakan adalah *CHELLAPILLA's Algorithm* yang bekerja dengan prinsip erosi dan dilasi ketika potongan-potongan piksel berukuran kecil lebih tipis daripada karakter. Tetapi algoritma ini memiliki kelemahan, yakni tidak dapat mengenali perbedaan antara karakter dan noise dengan lebar yang sama. Tingkat akurasi segmentasi yang dicapai pada sistem ini adalah 41,13%.

Berdasarkan masalah tersebut, pada Tugas Akhir ini penulis bermaksud melakukan segmentasi CAPTCHA berbasis teks berjenis Gimpy yang diimplementasikan pada *website* www.hotmail.com. Selain itu, pada penelitian sebelumnya sistem hanya dibuat hingga proses segmentasi, sehingga penulis mencoba untuk membuat sistem hingga proses pengenalan karakter. Dengan adanya sistem penguraian CAPTCHA ini, diharapkan CAPTCHA semakin berkembang dengan tingkat keamanan yang lebih tinggi untuk menghindari pihak-pihak yang berusaha mengakses *website* secara ilegal.

1.2 Tujuan Penelitian

Tujuan dalam pembuatan Proposal Tugas Akhir ini adalah seperti yang dijelaskan dibawah ini :

a. Menerapkan dan mensimulasikan algoritma pembaca CAPTCHA berbasis pengolahan citra.
b. Memecahkan tantangan CAPTCHA dengan jenis distorsi berupa *warping.*
c. Merancang sistem yang mampu menyelaraskan karakteristik karakter CAPTCHA yang beragam untuk dikenali.
d. Menguji performansi sistem melalui parameter tingkat akurasi dan waktu komputasi.
e. Melakukan perbandingan akurasi antara algoritma yang diterapkan pada sistem dengan hasil pembacaan *Optical Character Recognition* (OCR).

1.3 Rumusan Masalah

Rumusan masalah dalam pembuatan Proposal Tugas Akhir ini adalah seperti yang dijelaskan dibawah ini :

a. Bagaimana melakukan simulasi dan analisis CAPTCHA berbasis pengolahan citra menggunakan bahasa pemrograman MATLAB?

b. Bagaimana distorsi berupa dan *warping* mempengaruhi CAPTCHA dalam proses pemecahan tantangan?

c. Bagaimana melakukan *preprocessing* sebagai tahap awal perancangan sistem?

d. Bagaimana melakukan segmentasi teks menjadi beberapa karakter tunggal, sehingga mempermudah dalam proses identifikasi karakter?

e. Bagaimana melakukan proses klasifikasi karakter pada CAPTCHA ke dalam karakter yang benar?

f. Bagaimana mengukur kualitas keamanan dan ketahanan CAPTCHA dengan parameter tingkat akurasi dan waktu komputasi?

1.4 Batasan Masalah

Tugas akhir ini membatasi permasalahan pada poin-poin berikut :

a. CAPTCHA yang dikenali memiliki bentuk teks dan memiliki jenis *gimpy* yang diunduh dari halaman web www.hotmail.com.

b. CAPTCHA ditunjukkan melalui gambar dengan format .bmp.

c. Sistem yang dirancang bersifat *non-realtime*.

d. Bahasa Pemrograman yang digunakan adalah MATLAB 2009a.

1.5 Metode Penelitian

Penilitian ini dilakukan dengan metode-metode sebagai berikut :

a. Melakukan studi literatur

Pada tahap ini dilakukan dengan mempelajari permasalahan yang berkaitan dengan gimpy CAPTCHA, algoritma segmentasi dan proses pengenalan captcha. Proses pembelajaran ini dilakukan dengan kajian berbagai sumber pustaka baik berupa buku, jurnal ilmiah, maupun media elektronik.

b. Analisis dan simulasi

Pada tahap ini, akan dilakukan analisis untuk mengenali dan membaca sebuah CAPTCHA berjenis teks, dari mulai proses *preprocessing*, algoritma segmentasi yang digunakan, sampai pada proses pengenalan CAPTCHA itu sendiri. Kemudian mampu mensimulasikannya dengan program komputer.

c. Konsultasi dengan Dosen Pembimbing

Konsultasi dengan dosen pembimbing diperlukan untuk mengkaji dan merumuskan metode yang tepat untuk diimplementasikan dalam sistem sehingga hasil keluaran menjadi maksimal.

d. Analisis Kinerja Sistem

Tahap akhir dari penelitian tugas akhir ini adalah menganalisis data yang telah didapat pada tahap-tahap sebelumnya dan mampu menarik kesimpulan dan membandingkan hasilnya dengan jurnal yang mungkin telah ada sebelumnya.

1.6 Sistematika Penulisan

Tugas akhir ini dibagi dalam beberapa topik bahasan yang disusun secara sistematis sebagai berikut :

BAB I PENDAHULUAN

Bab ini menguraikan latar belakang penulisan, tujuan penulisan, rumusan masalah, batasan masalah, metodologi penulisan, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisi teori-teori yang mendukung dan mendasari penulisan tugas akhir ini, yaitu pengolahan citra *digital*, *preprocessing*, normalisasi karakter, metode segmentasi, serta pengenalan karakter tulisan pada gambar.

BAB III PERANCANGAN SISTEM DAN SIMULASI

Bab ini menjelaskan tentang proses perancangan yang meliputi *preprocessing*, segmentasi, normalisasi, ekstraksi ciri dan klasifkasi data uji CAPTCHA.

BAB IV ANALISIS DAN HASIL SIMULASI

Bab ini berisi analisa terhadap hasil yang diperoleh dari tahap perancangan sistem dan simulasi.

BAB V KESIMPULAN DAN SARAN

Bab ini berisikan kesimpulan dari analisa yang telah dilakukan dan saran untuk pengembangan ke depannya.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari Tugas Akhir yang berjudul Simulasi dan Analisis Penguraian Gimpy CAPTCHA Menggunakan Pengolahan Citra dan Segmentasi Huruf :

1. Telah berhasil dirancang suatu simulasi sistem penguraian CAPTCHA berbasis teks berjenis Gimpy yang diimplementasikan pada *website* www.hotmail.com.
2. Karakter pada CAPTCHA dapat dengan menerapkan prinsip pelabelan. Melalui pengaturan *threshold* dari `bwareaopen` sebesar 20 sistem akan membuang kumpulan piksel yang merupakan bagian dari karakter yang terpisah sehingga karakter akan tetap dipertahankan bentuk dasarnya. Dari hasil analisis, diperoleh tingkat akurasi segmentasi CAPTCHA terbaik mencapai 79,167% dengan 95 buah CAPTCHA tersegmentasi sempurna dan 25 buah CAPTCHA gagal disegmentasi.
3. Proses normalisasi karakter dalam rangka mengatasi keanekaragaman ketebalan, bentuk dan kemiringan dilakukan dengan operasi morfologi dan rotasi. Proses *thinning* dilakukan untuk menyamakan karakter agar memiliki ketebalan 1 piksel, rotasi dilakukan untuk mengatasi kemiringan karakter agar tegak dan dilasi dilakukan agar mengurangi *error* pada saat klasifikasi karakter.
4. Penambahan jumlah data ciri dapat mempermudah proses pengenalan karakter yang beraneka ragam sehingga meningkatkan tingkat akurasi, baik akurasi karakter ataupun akurasi *file* CAPTCHA.
5. Tingkat akurasi *file* dan akurasi karakter berhasil mencapai angka terbaik, yakni 12,5% untuk akurasi total *file* dan 70,76% untuk akurasi total karakter menggunakan skema yang sama, yakni metode ekstraksi *square* dan *resizing* citra karakter ke 24x24 piksel.

6. Metode ekstraksi *square* menghasilkan nilai akurasi yang lebih baik dibandingkan metode ekstraksi *sum* secara keseluruhan. Sedangkan metode *sum* unggul dalam waktu komputasi.
7. Penskalaan ulang *(resizing)* citra karakter mempengaruhi tingkat akurasi dan waktu komputasi. Secara keseluruhan, penskalaan dengan nilai yang lebih besar (24x24 piksel) menghasilkan nilai akurasi yang lebih tinggi
8. Sistem yang dirancang sudah memiliki akurasi yang lebih baik dalam mengenali karakter CAPTCHA dibandingkan dengan OCR.
9. Proses normalisasi membutuhkan waktu yang paling lama dibandingkan dengan proses lainnya, yakni *preprocessing*, segmentasi dan klasifikasi.
10. Waktu komputasi tercepat adalah 0,13929 detik yang terjadi pada skema *sum* dengan *resizing* ke 15x15 piksel, dimana diantara keseluruhan proses.
11. Secara keseluruhan, sistem dengan menggunakan metode ekstraksi ciri *square* dan ukuran penskalaan ke 24x24 piksel adalah skema terbaik sistem. Walaupun waktu komputasi tidak menghasilkan yang paling tercepat, selisih waktu komputasi masih dapat ditoleransi karena sistem menghasilkan selisih nilai tingkat akurasi yang signifikan.

5.2 Saran

Pengembangan lebih lanjut yang dapat dilakukan terhadap tugas akhir ini adalah sebagai berikut :

1. Peningkatan akurasi, baik akurasi segmentasi, *file*, ataupun karakter CAPTCHA menggunakan algoritma yang lebih baik dibandingkan sistem yang diimplementasikan pada tugas akhir ini.
2. Menampilkan nilai akurasi yang diperoleh dan verifikasi CAPTCHA terdeteksi benar atau tidak pada GUI *(Graphical Interface Unit)* yang dirancang.
3. Mengimplementasikan sistem penguraian Gimpy CAPTCHA ini pada bahasa pemrograman lain seperti C, java dan sebagainya.
4. Melakukan pengecekan keamanan untuk Gimpy CAPTCHA yang diimplementasikan pada *website* lainnya, seperti www.yahoo.com dan www.gmail.com.
5. Sistem dibuat secara *real-time*

DAFTAR PUSTAKA

[1] Abdillah, Muhammad Fikri. 2013. Analisis Segmentasi Citra Menggunakan Active Contour Model Pada Aplikasi Rambu Lalu Lintas. Bandung : Universitas Telkom.

[2] Anugroho Prasetyo dkk. 2010. Klasifikasi *Email* SPAM dengan Metode *Naive Bayes Classifier* Menggunakan *Java Programming*. Surabaya : Politeknik Elektronika Negeri Surabaya.

[3] Bursztein, Elie et al. 2011. *Text-Based CAPTCHA Strengths and Weaknesses*. ACM Computer and Communication Security.

[4] Caine, Allan dan Urs Hengartner. 2004. *The AI Hardness of CAPTCHAs does not imply Robust Network Security*. Canada : University Of Waterloo.

[5] Chandavale, A. A et al. 2009. *Algorithm To Break Visual CAPTCHA*. Second International Conference on Emerging Trends in Engineering and Technology

[6] Chandavale, A. A et al. 2009. *Reduced Process Thinning Algorithm for CAPTCHA Strength Measurement*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[7] Fiot, Jean-Baptiste et al. 2009. *The Captchacker Project*. Paris : ENS Cachan.

[8] Gao, Haichang et al. 2012. *Divide and Conquer : An Efficient Attack on Yahoo! CAPTCHA*. Software Engineering Institute Xidian University.

[9] Huang, Shih-Yu et al. 2008.*A Projection-based Segmentation Algorithm for Breaking MSN and Yahoo CAPTCHAs.* London.

[10] Louise Lam et al. 1992. *Thinning Metodologies-A Comprehensive Survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 14, No. 9.

[11] Megasari, Beby Nur. 2012. Pengenalan Kode Pos Berbasis Citra Tulisan Tangan yang Saling Bersentuhan dengan Algoritma Segmentasi *Background* dan *Foreground.* Bandung : Institut Teknologi Telkom.

[12] Mori, Greg. *Recognizing Objects in Adversarial Clutter Breaking a Visual CAPTCHA*. Berkeley : University Of California.

[13] Prasetyo, Eko. 2011. Pengolahan Citra Digital dan Aplikasinya Menggunakan MATLAB. Yogyakarta : Penerbit ANDI.

[14] Putra, Darma. 2009. Pengolahan Citra Digital. Yogyakarta : Penerbit ANDI.

[15] Santoso, Hendra. 2009. Captcha Sebagai Salah Satu Alat Untuk Mencegah Spambot. Malang : Malangkucecwara School of Economic.

[16] Saputra, Mohammad Adzif. 2013. Simulasi dan Analisis Segmentasi Citra Tulisan Tangan Angka yang Saling Bersentuhan Menggunakan Metode Zhang Suen. Bandung : IT Telkom.

[17] Sukamto, Rosa Ariani. 2008. Landasan Teori Thinning. Bandung : Institut Teknologi Bandung.

[18] Sutoyo, T dkk. 2009. Teori Pengolahan Citra Digital. Yogyakarta : Penerbit ANDI.

[19] Yan, J dan A. S. EI. Ahmad. 2007. *Breaking Visual CAPTCHAs with Naive Pattern Recognition Agorithms*. United Kingdom : School of Computing Science, Newcastle University.

[20] Yan, Jeff dan A. S. EI. Ahmad. 2009. *The Robustness of CAPTCHAs : A Security Engineering Perspective*. England : Newcastle University.