ABSTRACT

There is a lot of activity over the internet. Be it, posting a comment on someone's blog or making a Gmail account or booking a ticket online. But with that also comes the problem of spamming. "Completely Automated Public Turing test to tell Computers and Humans Apart" (CAPTCHA) [8] which are the twisted words that block the entries of bots on website. CAPTHCAs can effectively test if the user is human or machine. Hence it is of great importance that CAPTCHA is well checked for its vulnerability against such attacks. So, this paper presents this medium to check the strength of CAPTCHA against the written CAPTCHA cracking code. This can be used by the web developers implementing CAPTCHA, to check well in advance how secure is the CAPTCHA used in their software.

چکیده

فعالیت های زیادی از طریق اینترنت انجام می شود. ارسال نظر در وبلاگ شخصی یا ایجاد حساب Gmail یا رزرو آنلاین بلیط شامل فعالیت هایی است که شما از طریق اینترنت انجام می دهید. اما با این مسئله مشکل ارسال هرزنامه نیز وجود دارد. " آزمایش اتوماتیک عمومی برای تشخیص انسان از کامپیوتر" (captcha) که کلمات پیچیده ای هستند که از ورود ربات ها در وب سایت ها جلوگیری می کنند. CAPTHCA ها می توانند به طور موثری کاربر یا ماشین را آزمایش کنند. از این رو بسیار مهم است که CAPTCHA به خوبی از نظر آسیب پذیری در برابر چنین حملاتی بررسی شود. بنابراین این مقاله این رسانه را برای بررسی قدرت CAPTCHA در برابر کد ترک خورده CAPTCHA ارائه می دهد. توسعه دهنگان وب میتوانند CAPTCHA را پیاده سازی و استفاده کنند تا از قبل بدانند که CAPTCHA در نرم افزار آن ها چه قدر ایمن است.

General Terms

Security, Image Processing.

Keywords

CAPTCHA, OCR, Image Processing, MATLAB, Turing Test, CAPTCHA Cracking.

شرایط عمومی

امنیت ، پردازش تصویر.

کلید واژه ها

CAPTCHA، OCR ، پردازش تصویر ، MATLAB ، تست تورینگ ، CAPTCHA Cracking.

1. INTRODUCTION

In last few years, internet has witnessed brute force attacks as spammers developed bots to access websites and increase the load on the servers. This situation has caused new challenges and it demands the use of stronger CAPTCHAs. The plan is to crack these CAPTCHAs using "Image Processing". The future scope will include a testing module where it plans to test the complexity of CAPTCHA and hence assess the web developer in providing proper security measures for the website.

۱. مقدمه

در چند سال گذشته ، اینترنت با حملات بی رحمانه به عنوان اسپم ها ربات هایی را برای دسترسی به وب سایت ها و افزایش بار سرورها ، شاهد بوده است. این وضعیت چالش های جدیدی ایجاد کرده و استفاده از CAPTCHA های قوی تری را می طلبد. این برنامه برای شکستن این CAPTCHA ها با استفاده از "پردازش تصویر" است. هدف این است که در آینده یک ماژول آزمایش طراحی شودکه در آن قصد دارد پیچیدگی CAPTCHA را آزمایش کند و از این رو توسعه دهنده وب را در ارائه اقدامات امنیتی مناسب برای وب سایت ارزیابی کند.

2. RELATED WORK

CAPTCHA has been cracked by several organisations in the past with same motive to achieve higher security and stronger CAPTCHA sets [7]. The paper titled as „Stanford Researchers crack CAPTCHA code" by Todd Wasserman published by Stanford University has created DeCAPTCHA, software that makes CAPTCHA readable by computers by cleaning up the text and rendering them in legible letters and numbers. The tool decodes CAPTCHA most, but not all the time. The team was able to decode CAPTCHA up to 66 percentage accuracy. The paper titled as breaking an image based CAPTCHA by Michael Merler, Jacquilene Jacob published by Stanford University is on Vidoop CAPTCHA. It is a verification solution that uses images of the objects, and animals, people, instead of distorted text to distinguish a human from computer program [10]. What the authors underestimate that since a bot can try to access a service thousands of times a day, recognition rates which are considered quite low by the object recognition community.

۲. کار مرتبط

CAPTCHA در گذشته توسط چندین سازمان با انگیزه مشابه برای دستیابی به امنیت بالاتر و مجموعه های CAPTCHA قوی تر شکسته شده است. مقاله ای با عنوان "محققان استنفورد کد CAPTCHA را شکسته اند" توسط Todd Wasserman منتشر شده توسط دانشگاه استنفورد DeCAPTCHA را ساخته اند، نرم افزاری که با تمیز کردن متن و ارائه آنها با حروف و اعداد خوانا، CAPTCHA را برای رایانه قابل خواندن می کند. این ابزار بیشتر از همه CAPTCHA را رمزگشایی می کند ، اما نه همیشه. این تیم قادر به رمزگشایی CAPTCHA تا دقت ۶۶ درصد بود. مقاله ای با عنوان شکستن تصویر مبتنی بر CAPTCHA توسط Michael Merler، Jacquilene Jacob منتشر شده توسط دانشگاه استنفورد در Vidoop CAPTCHA است. این یک راه حل تأیید است که از تصاویر اشیا، و حیوانات ، مردم به جای متن تحریف شده برای تشخیص انسان از برنامه رایانه استفاده می کند. آنچه نویسندگان دست کم می گیرند این است که از آنجایی که یک ربات می تواند هزاران بار در روز سعی کند به یک سرویس دسترسی پیدا کند ، نرخ تشخیص آن توسط جامعه تشخیص اشیا بسیار کم در نظر گرفته می شود.

3. PROBLEM STATEMENT

The idea of the project is to break a text based CAPTCHA and to show that text based CAPTCHA are not highly secure. Currently the technology involved in cracking the CAPTCHA is not very accurate and has high processing time. But for a company implementing security using CAPTCHA they should know what level of security they are having by using a particular type of CAPTCHA. The software will take the screenshot of the website, crop the image to required area, process the image and return the answer. So basically the software will try to crack their CAPTCHA and then notify them the level of security they are having..

۳. بیان مسئله

ایده این پروژه این است که یک متن مبتنی بر CAPTCHA را شکسته و نشان دهد که متن مبتنی بر CAPTCHA از امنیت بالایی برخوردار نیست. در حال حاضر فناوری مربوط به شکستن CAPTCHA بسیار دقیق نیست و از زمان پردازش بالایی برخوردار است. اما برای شرکتی که امنیت را با استفاده از CAPTCHA

اجرا می کند ، آنها باید بدانند که با استفاده از نوع خاصی از CAPTCHA چه سطح امنیتی دارند. این نرم افزار تصویر صفحه وب را می گیرد ، تصویر را در قسمت مورد نیاز برش می دهد ، تصویر را پردازش می کند و پاسخ را برمی گرداند. بنابراین اساساً این نرم افزار سعی می کند CAPTCHA آنها را شکسته و سپس آنها را به سطح امنیتی که دارند مطلع سازد.

4. SOLUTION APPROACH AND METHODOLOGY

4.1 System Architectural Design

Chosen System Architecture
The system architecture designed for the proposed system is as follows: The components of the system architecture are described in detail as follows:
Data Collection: This component is primarily focused on acquisition of data which consists of collection of CAPTCHA images.
Decision making: This component deals by first takes the CAPTCHA image processes it and returns the text in it.

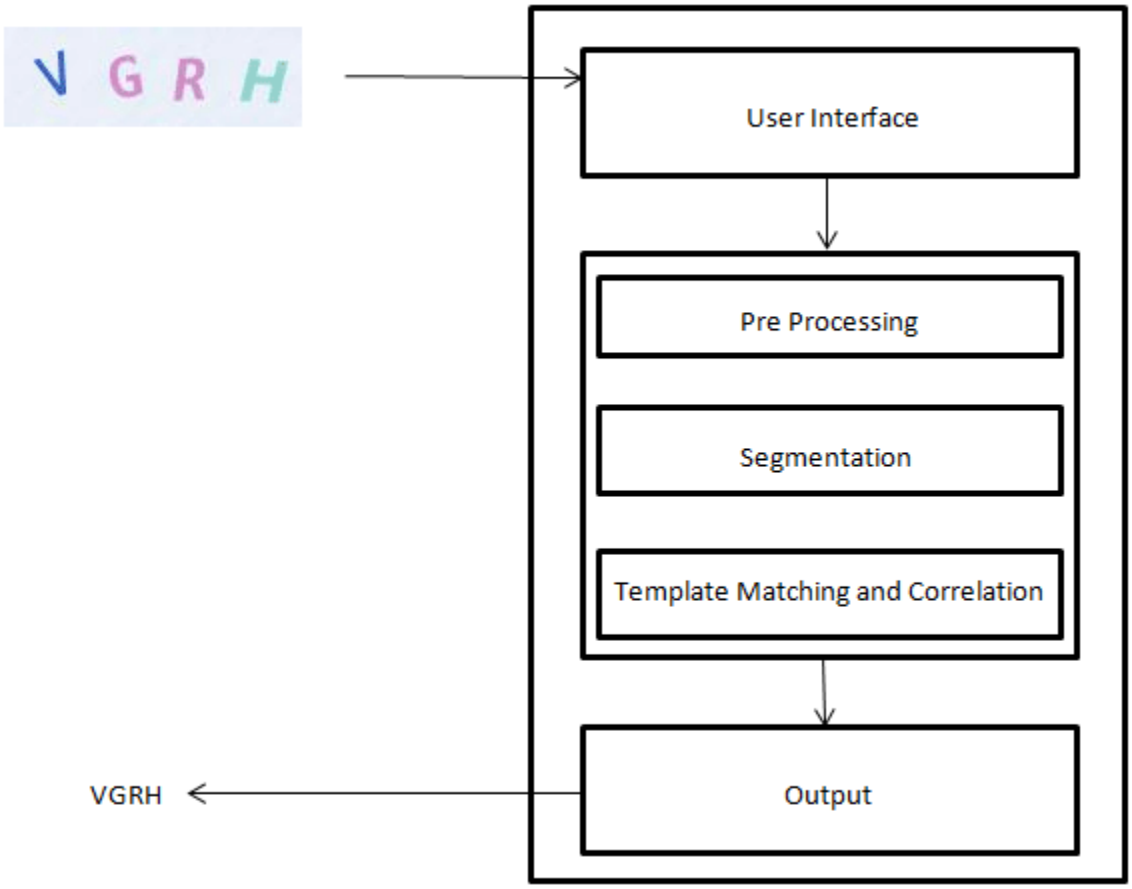


Figure 1: System Architecture

4.2 Detailed Description of Component

Data collection: This component primarily deals with the collection of the input data that is the list of all the CAPTCHA images. The images are then processed in the processing stage.
Data processing: The user should be feeding the link of the webpage for which the CAPTCHA needs to be cracked. The application will be able to extract CAPTCHA as an image. The application will process the image and recognize characters of the CAPTCHA. The application will provide the answer.

۴. رویکرد و روش حل

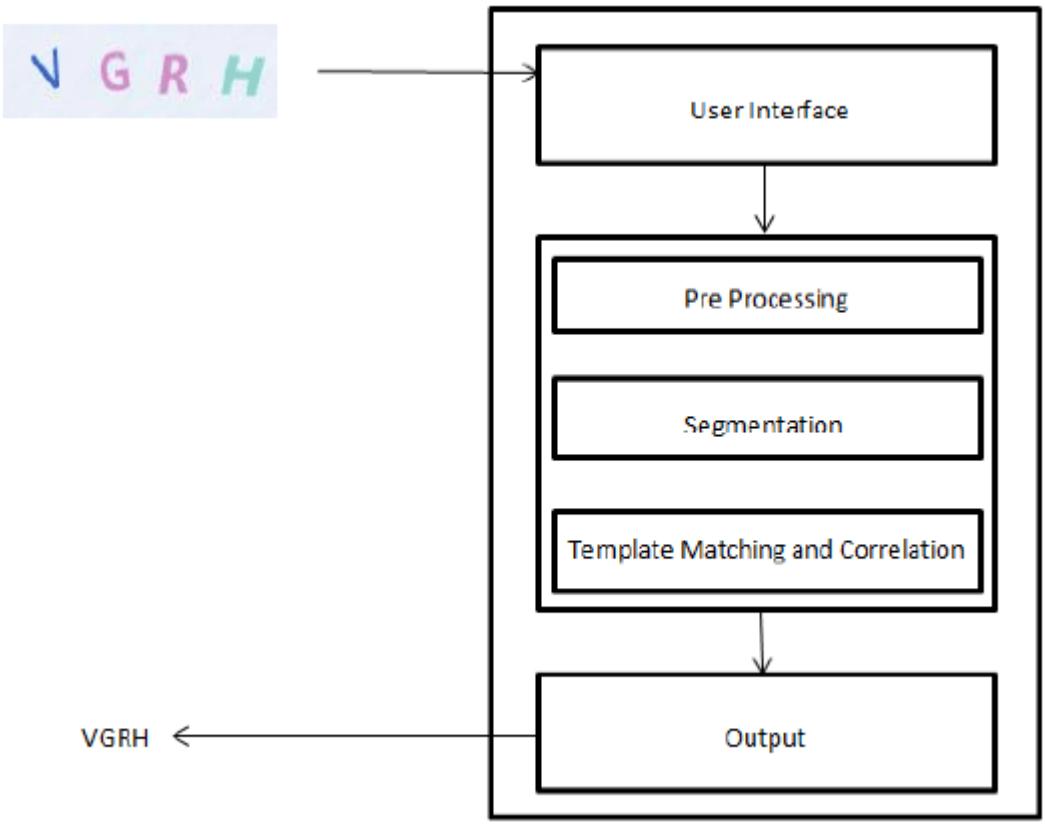
۴.۱ طراحی معماری سیستم

معماری سیستم برگزیده

معماری سیستم طراحی شده برای سیستم پیشنهادی به شرح زیر است: اجزای معماری سیستم به تفصیل به شرح زیر شرح داده شده است:

جمع آوری داده ها: این مولفه در درجه اول بر جمع آوری داده هایی متمرکز است که شامل جمع آوری تصاویر CAPTCHA است.

تصمیم گیری: این مولفه در وهله اول تصویر CAPTCHA را پردازش می کند و متن را در آن بر می‌گرداند.



شکل ۱: معماری سیستم

۴.۲ شرح دقیق جز مولفه

جمع آوری داده ها: این مولفه در درجه اول با جمع آوری داده های ورودی سروکار دارد که لیستی از تمام تصاویر CAPTCHA است. سپس تصاویر در مرحله پردازش پردازش می شوند.

پردازش داده ها: کاربر باید پیوند صفحه وب را که CAPTCHA برای آن شکسته است تغذیه کند. این برنامه قادر به استخراج CAPTCHA به عنوان تصویر خواهد بود. برنامه تصویر را پردازش می کند و شخصیت های CAPTCHA را تشخیص می دهد. برنامه جواب را آماده میکند.

4.3 Discussion of Alternative Designs

Alternative designs can be implemented for the proposed system. It can use the cut point detector to find all the **possible** cuts along which to segment CAPTCHA into individual characters and then slicer for getting some meaningful slices then scorer and arbiter for getting the text out of the CAPTCHA [7].

۴.۳ بحث در مورد طرح های جایگزین

طرحهای جایگزین را می توان برای سیستم پیشنهادی اجرا کرد. این می تواند از آشکارساز نقطه برش برای پیدا کردن تمام برش های ممکن استفاده کند که از طریق آن CAPTCHA تقسیم می شود به کاراکتر های جداگانه و سپس برش دهنده برای بدست آوردن برش‌های معنی‌دار از scorer و arbiter برای گرفتن متن از CAPTCHA مورد استفاده قرار میگیرند.

4.4 Component Diagram

Output interface: The final output will show the result which will be access granted or access denied
Files: Source files, executable files, database files.
Libraries: **MATLAB** libraries.

۴.۴ نمودار اجزا

رابط خروجی: خروجی نهایی نتیجه ای را نشان می دهد که اجازه دسترسی داده می شود یا دسترسی از آن سلب می شود.

پرونده ها: پرونده های منبع ، فایلهای اجرایی ، پرونده های پایگاه داده.

کتابخانه ها: کتابخانه های متلب.

4.5 Decision making

This component merely decides whether the given CAPTCHA is cracked or not and if it is cracked it is not secured and **hence** not suitable for the website. This component interacts with the software, the data processing unit and the user.

۴.۵ تصمیم گیری

این مولفه فقط تصمیم می گیرد که آیا CAPTCHA داده شده کرک شده است یا خیر و اگر کرک شده امن نیست و بنابراین برای وب سایت مناسب نیست. این مولفه با نرم افزار ، واحد پردازش داده و کاربر تعامل دارد.

4.6 Use Case Diagram

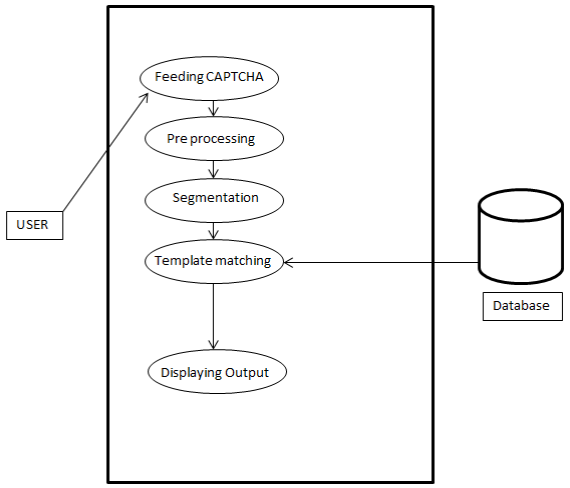
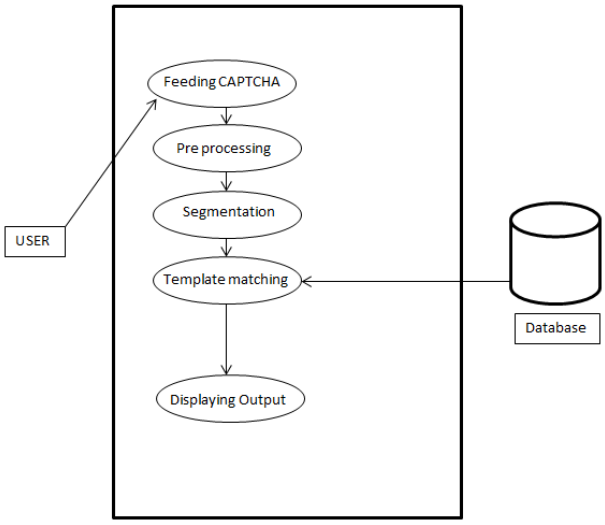


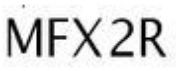
Figure 2: Use Case Diagram

۴.۶ استفاده از نمودار موردی



شکل ۲: استفاده از نمودار موردی

4.7 Algorithm to crack a type of CAPTCHA



Preprocessing

The CAPTCHA image is converted to gray scale image if it's a RGB image. After converting RGB image to gray scale image, the threshold method is performed. First it **calculates** the threshold value using predefined function graythresh(). This threshold value is used, where gray-levels below this threshold is said to be black and levels above are said to be white [1]. It performs binarization for further segmentation [6].

Figure 3: Preprocessed image

۴.۷ الگوریتم برای شکستن نوعی CAPTCHA

پیش پردازش

تصویر CAPTCHA در حالت RGB به تصویر خاکستری تبدیل می شود. پس از تبدیل تصویر RGB به تصویر در مقیاس خاکستری، روش threshold انجام می شود. ابتدا مقدار threshold را با استفاده از تابع از پیش تعیین شده greythresh() محاسبه می کند. این مقدار آستانه استفاده می شود ، جایی که گفته می شود سطح خاکستری زیر این آستانه سیاه است و گفته می شود که سطوح بالاتر سفید است. برای تقسیم بندی بیشتر باینری سازی را انجام می دهد.

شکل ۳: تصویر پیش پردازش شده

Segmentation

Segmentation is the isolation of characters or words. The majority of optical character recognition algorithms segment the words into isolated characters **which** are recognized individually as shown in figure 4. This segmentation is performed by isolating each connected component that is each connected white area. In matrix term each connected 1"s can be termed as one segment as shown in figure 5 [2].

Figure 4: Segmented character

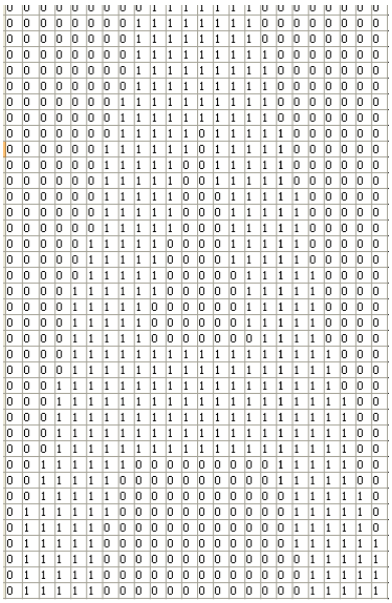
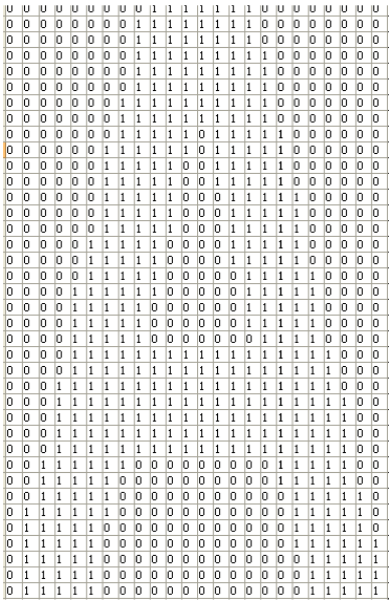


Figure 5: Segmented character in form of matrix

تقسیم بندی

تقسیم بندی برای ایزوله کردن کاراکتر ها یا کلمات است. اکثر الگوریتم های تشخیص کاراکتر نوری، کلمات را به کاراکتر جداگانه تقسیم می کنند که به صورت جداگانه شناسایی می شوند. همانطور که در شکل ۴ نشان داده شده است. این تقسیم بندی با جداسازی هر مولفه متصل که هر یک از مناطق سفید متصل است، انجام می شود. در اصطلاح ماتریس ، هر ۱ مولفه متصل را می توان به عنوان یک بخش تعریف کرد. همانطور که در شکل ۵ نشان داده شده است.

شکل ۴: کاراکتر تقسیم شده



شکل ۵: کاراکتر تقسیم شده به شکل ماتریس

Template-matching and correlation techniques

These techniques are different from the others in that no features are actually extracted. Instead, the matrix containing the image of the input character is directly matched with a set of characters in templates representing each possible class. The distance between the pattern and each character in template is computed, and the maximum correlation value obtained from segmented character and the characters in template is the character printed to output [2].

تکنیک های همسان سازی الگو و همبستگی

این تکنیک ها از این نظر متفاوت هستند که هیچ ویژگی در واقع استخراج نمی شود. در عوض ماتریس حاوی تصویر کاراکتر ورودی مستقیماً با مجموعه ای از کاراکتر ها در الگوهای نمایانگر هر کلاس ممکن مطابقت دارد. فاصله بین طرح و هر کاراکتر در الگو محاسبه می شود و حداکثر مقدار همبستگی حاصل از کاراکترهای تقسیم شده و کارکتر های موجود در الگو ، کاراکتر چاپ شده برای خروجی را به ما میدهد.

Cracking of second type of CAPTCHA

This type of CAPTCHA suffers from several weaknesses, be it fixed font face, fixed font size, no distortions, trivial background noise and it's easy to segment. Here the three-step algorithm is used to break the CAPTCHA. The image is preprocessed to remove noise using threshold method.

Thereafter a simple cleaning technique is used to clean the noise. Then the CAPTCHA is segmented using vertical projections and candidate split positions. Four classification methods have been implemented which are pixel counting, vertical projections, horizontal projections and template correlations. The system was trained on a sample of twenty CAPTCHAs to create thirty-six training templates one for each character (0-9 and A-Z). The following success rates have been achieved using the different classifiers: 8% pixel counting, vertical projections 97%, horizontal projections 100%, and template correlations 100%.

شکستن نوع دوم CAPTCHA

این نوع CAPTCHA از چندین نقطه ضعف رنج می برد ، چه صورت قلم ثابت ، چه اندازه قلم ثابت ، بدون تحریف ، نویز پس زمینه بی اهمیت و تقسیم بندی آن آسان است. در اینجا از الگوریتم سه مرحله ای برای

شکستن CAPTCHA استفاده می شود. تصویر برای حذف نویز با استفاده از روش threshold پیش پردازش می شود. پس از آن از یک روش تمیز کردن ساده برای از بین بردن نویز استفاده می شود. سپس CAPTCHA با استفاده از پیش بینی های عمودی و موقعیت های تقسیم انتخاب شده، تقسیم بندی می شود. چهار روش طبقه بندی اجرا شده است که شمارش پیکسل ، پیش بینی عمودی ، پیش بینی افقی و همبستگی الگو است. این سیستم بر روی نمونه ای از بیست CAPTCHA آموزش داده شد تا سی و شش الگوی آموزشی ایجاد کند که یکی برای هر کاراکتر (۰-۹ و A-Z) ایجاد شود. نرخ موفقیت زیر با استفاده از طبقه بندی کننده های مختلف بدست آمده است: ۸٪ شمارش پیکسل ، پیش بینی عمودی ۹۷٪ ، پیش بینی افقی ۱۰۰٪ و همبستگی الگو ۱۰۰٪.

Algorithm to crack this CAPTCHA

Steps involved are as follows.

Making of template

Firstly, save the CAPTCHA images with file name as its output. Load all the **images** from the training directory. Then perform preprocessing and segmentation as described below. Now the segmented characters of training CAPTCHA are mapped with corresponding characters of file name and saved into template database.

الگوریتم برای شکستن این CAPTCHA

مراحل انجام شده به شرح زیر است.

ساخت الگو

ابتدا تصاویر CAPTCHA را با نام پرونده به عنوان خروجی ذخیره کنید. تمام تصاویر را از فهرست آموزش بارگیری کنید. سپس پیش پردازش و تقسیم بندی را به شرح زیر انجام دهید. اکنون کاراکتر های تقسیم شده آموزش CAPTCHA با کاراکتر های مربوط به نام فایل ترسیم شده و در پایگاه داده الگو ذخیره مطابقت داده می شوند.

Preprocessing

The CAPTCHA is fed in the CAPTCHA cracker software and converted to gray scale as show in figure 6 using MATLAB inbuilt function rgb2gray(). Further the CAPTCHA image is threshold as show in figure 6 and image is cleaned of noise present in it as shown in figure 9. The final output of this step is preprocessed image.

Fig 6: Original Image

Fig 7: Grey Scale Image

Fig 8: Threshold Image

Fig 9: Further Cleaned Image

پیش پردازش

CAPTCHA در نرم افزار کرک کننده CAPTCHA تغذیه می شود و همانطور که در شکل ۶ با استفاده از تابع داخلی rgb2gray() در متلب که نشان داده شده است، به مقیاس خاکستری تبدیل می شود. علاوه بر این، تصویر CAPTCHA همانطور که در شکل ۶ نشان داده شده است، threshold شده است و تصویر از نویز

موجود در آن پاک می شود. همانطور که در شکل ۹ نشان داده شده است. خروجی نهایی این مرحله تصویر پیش پردازش شده است.

شکل ۶: تصویر اصلی

شکل ۷: تصویر در مقیاس خاکستری

شکل ۸: تصویر Threshold شده

شکل ۹: تصویر تمیز تر (بدون نویز تر)

Segment

The preprocessed CAPTCHA block is segmented into individual characters as shown in figure 10. Detailed explanation of this step is as follows. Firstly, the size of preprocessed image is obtained, then every individual column is scanned till the data is found. Then the obtained character is cropped. Twenty rows and columns are padded with 0's as shown in fig 11.

Fig 10: Segmented Image

Fig 11: Padded Image

بخش

بلوک CAPTCHA پیش پردازش شده به کاراکتر های جداگانه تقسیم می شود همانطور که در شکل ۱۰ نشان داده شده است. توضیحات دقیق این مرحله به شرح زیر است. ابتدا اندازه تصویر پیش پردازش شده بدست می آید ، سپس هر ستون جداگانه اسکن می شود تا داده ها پیدا شود. سپس کاراکتر بدست آمده بریده می شود. همانطور که در شکل ۱۱ نشان داده شده است ، ۲۰ ردیف و ستون با ۰ پر می شوند.

شکل ۱۰: تصویر تقسیم شده

شکل ۱۱: تصویر پر شده

Classify

Firstly the templates are loaded then it can use four types of classification techniques as follows:

طبقه بندی کردن

ابتدا الگوها بارگذاری می شوند و سپس می توانند از چهار نوع روش طبقه بندی به شرح زیر استفاده کنند:

Pixel Count

First, the vertical segmentation divides the characters into 5 segments. Next each segment is scanned to get the number of foreground pixels in it. Then, the pixel count obtained in the previous step is used to look up the mapping table [11.. Finally, it gives the output with success rate of 8%.

شمارش پیکسل

ابتدا تقسیم بندی عمودی کاراکتر ها را به ۵ بخش تقسیم می کند. بعد هر بخش اسکن می شود تا تعداد پیکسل های پیش زمینه را بدست آورد. سپس ، از تعداد پیکسل بدست آمده در مرحله قبل برای جستجوی جدول نگاشت استفاده می شود. سرانجام با موفقیت ۸٪ خروجی می دهد.

Vertical Projections

Vertical projection is applied to a segmented image, each of which contains one character. The process of vertical projection starts by mapping the image histogram to that of the template vertical histogram which finally has more mapping involved to it is the output. Finally, it gives the output with success rate of 95% [3].

پیش بینی های عمودی

تصویر عمودی بر روی یک تصویر تقسیم بندی شده اعمال می شود ، که هر یک شامل یک کاراکتر است. فرآیند تصویر عمودی با نگاشت هیستوگرام تصویر بر روی هیستوگرام عمودی الگو آغاز می شود که درنهایت نگاشت بیشتری در خروجی آن نقش دارد. سرانجام با موفقیت ۹۵٪ خروجی می دهد.

Horizontal Projections

Horizontal projection is applied to a segmented image, each of which contains one character. The process of vertical projection starts by mapping the image histogram to that of the template horizontal histogram which finally has more mapping involved to it is the output. Finally, it gives the output with success rate of 100%.

تصویر افقی

تصویر افقی به یک تصویر تقسیم شده اعمال می شود ، که هر یک شامل یک کاراکتر است. فرآیند تصویر عمودی با نگاشت هیستوگرام تصویر بر روی هیستوگرام افقی الگو آغاز می شود که در نهایت نگاشت بیشتری در خروجی آن نقش دارد. سرانجام با موفقیت ۱۰۰٪ خروجی می دهد.

Template Correlations

The matrix containing the image of the input character is directly matched with a set of characters in templates representing each possible class. The distance between the pattern and each character in template is computed, and the maximum correlation value obtained from segmented character and the characters in template is the character printed to output. Finally it gives the output with success rate of 100%.

همبستگی های الگو

ماتریس حاوی تصویر کاراکتر ورودی مستقیماً با مجموعه ای از نویسه ها در الگوهای نمایانگر هر کلاس ممکن مطابقت دارد.

فاصله بین مدل و هر کاراکتر در الگو محاسبه می شود و حداکثر مقدار همبستگی بدست آمده از کاراکتر تقسیم شده و کاراکتر های موجود در الگو ، کاراکتر چاپ شده برای خروجی است. سرانجام با موفقیت ۱۰۰٪ خروجی می دهد.

5. CONCLUSION

As we know increase in number of CAPTCHA usage in every web services. As CAPTCHA is termed as secure to prevent DOS attacks as it avoids running of automated scripts.

Finally the conclusion by cracking the CAPTCHA is that it is not completely secure. This paper shows the technique to crack CAPTCHA so that one can implant CAPTCHA cracker by giving CAPTCHA image as an input and retrieving CAPTCHA text as output.

۵. نتیجه گیری

همانطور که می دانیم افزایش تعداد CAPTCHA در هر وب سرویس افزایش یافته است. از آنجا که CAPTCHA برای جلوگیری از حملات DOS ایمن شناخته می شود زیرا از اجرای اسکریپت های خودکار جلوگیری می کند.
سرانجام نتیجه گیری با شکستن CAPTCHA این است که کاملاً ایمن نیست. این مقاله تکنیک شکستن CAPTCHA را نشان می دهد به طوری که می توان با دادن تصویر CAPTCHA به عنوان ورودی و بازیابی متن CAPTCHA به عنوان خروجی ، کرکر CAPTCHA را به عنوان خروجی دریافت کرد.

6. FUTURE SCOPE

An automatic software can be made which will auto detect the CAPTCHA in the website and will feed the CAPTCHA to the website automatically. The solutions can also be provided to them about which security levels they can include in their CAPTCHA.Automated learning can also be given to the CAPTCHA cracker which is supported by human so that multiple training sets can be created.

۶. اهداف آینده

یک نرم افزار اتوماتیک می تواند ساخته شود که به صورت خودکار CAPTCHA را در وب سایت شناسایی کند و به صورت خودکار CAPTCHA را به وب سایت تغذیه کند. همچنین می توان برای آنها در مورد سطح امنیتی که می توانند در CAPTCHA خود قرار دهند راه حلهایی ارائه شود. همچنین می توان به کرکر CAPTCHA که توسط انسان پشتیبانی می شود ، یادگیری خودکار داده شود تا مجموعه های آموزشی متعددی ایجاد شود.