Information Security

CECS 579



**Documentation**

**Hack CAPTCHA**

**Member:**

**Karishma Mehta (013656966)**

**Professor:**

**Mehdrad Aliasgiri**

1. Abstract

Hack CAPTCHA is a project aimed at cracking CAPTCHA images. CAPTCHA are termed as “Completely Automated Public Turing Test To Tell Computers and Human Apart by Luis von Ahn, Manuel Blum, Nicholas Hopper and John Langford of Carnegie Mellon University in 2000[1]. The main purpose of this program is to protect websites from bots accessing it. It generates few grading tests that only humans can pass and cannot be passed by any computer.

Few advantages of CAPTCHA are as follows:

1. It helps to prevent from comment spam in blogs

2. Used while registering to a website

3. Helps in preventing dictionary attacks

In the context of CAPTCHA, the aim of this project is to break CAPTCHA or in other terms to “HACK-CAPTCHA”. It can be thought of taking an image and making it pass through an algorithm which will output the text of the image. We can also calculate failure probability and success probability of hacking CAPTCHA.

There are algorithms available in Machine Learning which can help us to achieve this. The various phases of algorithm in Machine learning are composed of:

1. Pre-processing
2. Segmentation
3. Feature Extraction
4. Character Recognition

Various algorithm like k-nearest neighbor, neural network algorithm, Vector Space algorithm are based on the above steps.

Examples of CAPTCHA:

 

**Input Image Output Image**

In the above input image we can see lot of noise, distortion which is first removed from it. Removing background noise leaves us with following output image

**Artificial Intelligence and Vector Space Image Recognition**

Image processing is considered to be the biggest success in the field of Artificial Intelligence. Any particular individual studying AI is bound to come across using “Neural Network” . There is hardly any success with neural network to learn how to identify characters. But later the technique of Vector Space Image Recognition was introduced which provided better results as per the studies obtained from few research papers

There are few advantages of vector space image algorithms which include

* Extensive training iterations not needed
* It is not possible for them to become overtrained
* It is possible to add / remove incorrect data whenever needed
* Graded results help us to see top –x results
* If you are not able to recognize something, just simply add it and you can recognize instantly

On the other side these algorithms are much slower than neural networks. The other problem with these algorithms is that they don’t find their own way of solving the problem. This analysis was very well provided in a paper on “Basic Vector Space Search Engine Technology” by Vidiot[2]

**Building Training Set**

For this step we need to have collection of images with which we want to compare our images. This particular training set can be used by any AI algorithm which we have built. Inappropriate use of datasets can also break the system. It is very much important to select appropriate dataset in order to succeed. A good thing is to include all those CAPTCHA images of similar type. Some images are too curved, some are little. Just as to keep the complexities low, I have used less complicated CAPTCHA’s

**The final step**

The final step would be to be put everything together. We need to have 4 components:

1. Text extracter
2. Letter extracter
3. Recognition technique
4. Training set

**Technology Used:**

Since this is the topic of AI, finding patterns in data, Python works lot well with it. It has various libraries like PIL(Python Image Library), OpenCV. It also has numpy framework to work with machine learning algorithm

Python openCV with pytesser can break a CAPTCHA in just 26 lines of code!! I am not taking up this approach, because it simply takes in an image file and returns the data in string format

I would actually implement an algorithm to break the system.

For GUI, I am using the PyGTK library

**References**

[1]CAPTCHA Definition - <http://www.captcha.net/>

[2]Basic Vector Space Search Engine Theory http://la2600.org/talks/files/20040102/Vector\_Space\_Search\_Engine\_Theory.pdf