網頁圖形驗證碼識別與破解實驗

103318098 黃柏翔

前言

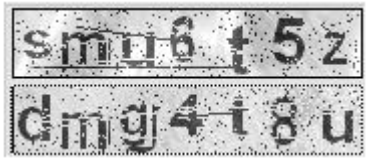
- 隨著電腦運用的普及與網際網路的蓬勃發展，越來越多的服務皆採用電子化(e化)作業。
- 網路上來自不同社會階級的使用者人數增加。有些是名人，有些是普通人，當然壞人也確實存在。
- 近年來電腦病毒、駭客入侵及個人資料外洩等資安事件發生越來越頻繁。
- 資訊便利也會帶來資訊安全問題，因此必須做好資訊安全防護措施。

研究動機與目的

- 很多網路上的攻擊，尤其是重複性高的動作，往往都是機器人在執行。
- 為了防止網路機器人攻擊*(像是暴力破解密碼)*或是操作大量重複性的動作*(像是張貼垃圾廣告訊息)*，網頁上的表單常常會採用圖形驗證碼來保護，識別使用者是『人類』還是『機器人』。
- 使用者每次操作都需要輸入圖形驗證碼，偶爾操作還好，若要常常操作都要輸入圖形驗證碼會非常的不方便。
- 以下的實驗會嘗試自動識別並嘗試輸入正確的圖形驗證碼。

驗證碼介紹

- 驗證碼，就是將一串隨機產生的數字或符號，生成一幅圖片，圖片裡加上一些干擾像素，以防止光學字元識別(OCR)。
- 光學字元識別(OCR)是指對文字資料的影像檔案進行分析識別處理，取得文字及版面資訊的過程。
- 由使用者肉眼識別其中的驗證碼字元，輸入表單提交給網站做驗證，驗證成功後才能使用該網站的功能。



驗證碼種類介紹

- fig 1，圖片背景和數字都使用單一的顏色，而且字元整齊，位置統一。
- fig 2，加入背景色和雜訊干擾線條，但是字元還算整齊，顏色也相同。
- fig 3，除了背景色和干擾像素，字元的顏色也有變化，並且顏色各不相同。
- fig 4，在文字圖片上加入了兩條直線干擾。

fig 1

fig 2

fig 3

fig 4

基本處理步驟介紹

- Step 1 取出字模

首先取出字模就是將要破解的圖形驗證碼先抓取回來，而取得的字模圖片必須要包含所有會出現的文字，例如 0~9 a~z 的字元圖片，當有了字模後就能夠將字模進行二值化。

基本步驟介紹

- Step 2 二值化

二值化就是將數字字模轉換成 0 與 1 的結果，將圖片上數字的部分用 1 替換而 0 則代表背景，例如我有一張數字 3 的圖片，在經過二值化後就會變成以下結果。

3

基本步驟介紹

- Step 3 計算特徵

取得驗證碼二值化的值之後，因為驗證碼可能包含干擾元素，必須要先去除干擾元素後將圖片二值化取得特徵。

像是背景雜訊、干擾線等等，都會在這個步驟做過濾。

基本步驟介紹

- Step 4 對照樣本

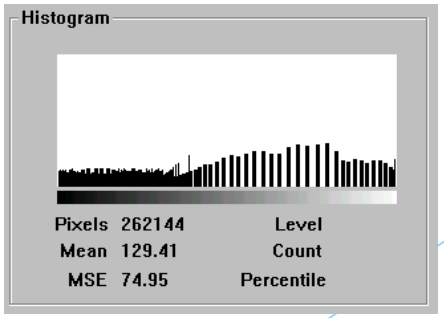
最後的步驟就是要將第三步驟二值化的值拿去比對我們的樣本庫，通常在比對的時候一定會產生誤差值，例如以下轉換後的二進值，可以看到以上二進值紅色的 1 的部分就是所謂的噪點，我們可以透過一些演算法解決這個問題，盡量找出與樣本中最相近的結果。

3

演算法介紹

灰度直方圖計算(histogram)

- 灰度直方圖（histogram）是灰度級的函數，它表示圖象中具有每種灰度級的像素的個數，反映圖象中每種灰度出現的頻率。
- 灰度直方圖的橫坐標是灰度級，縱坐標是該灰度級出現的頻率。
- 要計算圖像的相似度，就要找出圖像的特徵，常用的圖像特徵有顏色特徵、紋理特徵、形狀特徵和空間關係特徵等等。
- 直方圖能夠描述一幅圖像中顏色的分布，而且容易理解和實現，所以大部分圖像的相似度計算均使用它。



演算法介紹

灰度直方圖計算(histogram)

- 得到圖形的直方圖後，圖像的相似度計算公式如下：

$$Sim(G,S)=\frac{1}{N}\sum_{i=1}^{N}\left[1-\frac{|g_i-s_i|}{Max(g_i,s_i)}\right]$$

- 其中，G、S為直方圖，N為顏色空間樣點數
- gi和si分別為兩張圖片的直方圖縱坐標值，即樣點數。
- Sim(G,S) 計算結果越大，表示兩張圖片越相似。

演算法介紹 感知哈希算法 (Perceptual Hash Algorithm)

- 感知哈希算法(Perceptual Hash Algorithm)會為每張圖片生成一個指紋(字串格式)，當兩張圖片的指紋越相似，就說明兩張圖片越相似。
- Google 相似圖片搜索也是使用『感知哈希算法』達成。

演算法介紹 感知哈希算法 (Perceptual Hash Algorithm)

計算出指紋的步驟為：

- 1. 縮小圖片尺寸

將圖片縮小到8x8的尺寸，總共64個像素。 這一步的作用是去除各種圖片尺寸和圖片比例的差異，只保留結構、明暗等基本資訊。

- 2. 轉為灰度圖片

將縮小後的圖片，轉為64級灰度圖片。

- 3. 計算灰度平均值

計算圖片中所有像素的灰度平均值

演算法介紹 感知哈希算法 (Perceptual Hash Algorithm)

- 4. 比較像素的灰度

將每個像素的灰度與平均值進行比較，如果大於或等於平均值記為1，小於平均值記為0。

- 5. 計算哈希值

將上一步的比較結果組合在一起，就構成了一個64位的二進制整數，這就是這張圖片的指紋。

- 6. 對比圖片指紋

得到圖片的指紋後，就可以對比不同的圖片的指紋，計算出64位中有多少位是不一樣的。如果不相同的數據位數不超過5，就說明兩張圖片很相似，如果大於10，則說明它們是兩張不同的圖片。

實驗操作


Captcha Cracker BETA v2014.12.31 By MapleHuang(stan60250@gmail.com)
/index.do?thetime=1420009260143
網頁/流程控制
圖形處理
分割/樣本工具/識別演算
樣本管理
除錯
自動登入設定
帳號：103318098
密碼：********
次數：1
登出
開始自動登入
網頁控制
[網頁]前往
逾時(秒)：30
[網頁]停止
[網頁]重整
使用者帳號：
(Account)
使用者密碼：
(Password)
驗證碼：
(Code)
apyl
請輸入驗證碼：
(Keyin Code)
登入 (Login)

實驗結果

- 採用灰度直方圖計算法，圖形辨識正確率幾乎高達100%
- 採用感知哈希算法，圖形辨識正確率幾乎高達100%
- 這代表網站的圖形驗證碼過於簡單，網站有被攻擊或是大量重複操作的可能性。

總結

- 如果真的做出可以完美破解圖形驗證碼的工具，就代表網頁本身圖形驗證碼功能不夠完善，也不夠安全。
- 過於簡單的圖形驗證碼沒辦法正確區分『人類』與『機器人』。
- 可以根據實驗結果改良網頁圖形驗證碼功能，像是加強圖形驗證碼的辨識難度，或是更換成其他的驗證方式，讓網站更加安全。