OCR

M – AI – Machine Learning – Part I

1. General

OCR, literally: Optical Character Recognition is a mechanism used to digitize handwritten or machine-written characters. These characters can be Latin letters, numbers, or other symbols like punctuation, by example. Of course, this is a general concept, it can be extended to Cyrillic, Chinese, Arabic, or any other type of character.

This method is very widely used, and is now in the everyday life: on your computers, your smartphones, etc. You can use it to write with a tablet like on a paper, and translate it to plain text, much easier to use in your software. Or, you can digitalize scanned books to improve a further numeric reading in your e-paper tool: by example, Google is investing a lot of budget in this technology in order to build a huge database of books.

Machine Learning is a prerequisite to make such an adaptive tool: you can never presume how a human write, nor describe to your computer a global concept, merely as simple as the letter “A”. Use the power of supervised learning, and craft a powerful OCR!

2. Goals

For this project, you will have to produce an OCR tool able to recognize handwritten characters. We expect you to pass two main tests:

1. Your OCR must be able to recognize individual characters. The charset is the printable part of the ASCII table (refer to the example dataset).
2. Your OCR must be able to digitalize entire handwritten texts.

We do not give you any method, it is up to you to study and choose a proper algorithm. You will have to explain and justify your choice. Produce every document and compiled result you have! The design is an important part of the work.

Moreover, you will have to craft a simple and usable test GUI, so your live demonstration will be fancier than a terminal!

3. STEPS AND DATA SETS

For the purpose of this project, you will have to provide some data handcrafted by yourself. It is mandatory! Everything must be produced by hand, no digital font will be accepted. Use a scanner, a graphic tablet, a smartphone, etc.

Each member of a group has to do it!

Upload it in the repository of which you should have received information by e-mail.

An example data set is available, you will find it with this document. Refer to it for file names and co.

3.1. FIRST STEP: CHARACTERS

Provide a set of images with these characteristics:

- **Size** : 50 x 50 px, 32 bits
- **Name** : `<login>-<char>.bmp`
  Replace <char> by the char name, <login> by your login.
- **Archive** : Compress it in `<login>_step1.zip`
- **Directory** : `/step1`

Characters corresponds to the printable part of the ASCII table.

3.2. SECOND STEP: TEXT

Provide a text of your choice. It must fit on the image:

- **Size** : 831 x 594 px, 32 bits
- **Name** : `<login>-text.bmp`
  Replace <login> by your login.
- **Directory** : `/step2`

3.3. BONUS STEP: CAPTCHA!

Ultra bonus! Make a tool able to solve some re-captchas, and you will be rewarded!

You don’t have to provide images. The ones in the example dataset are provided for illustration.

http://www.google.com/recaptcha

4. AUTHORIZATIONS AND RESTRICTIONS

You are allowed to use every library which can help you.

Of course, you are not allowed to directly use an OCR tool! You will have to design yourself your algorithm.