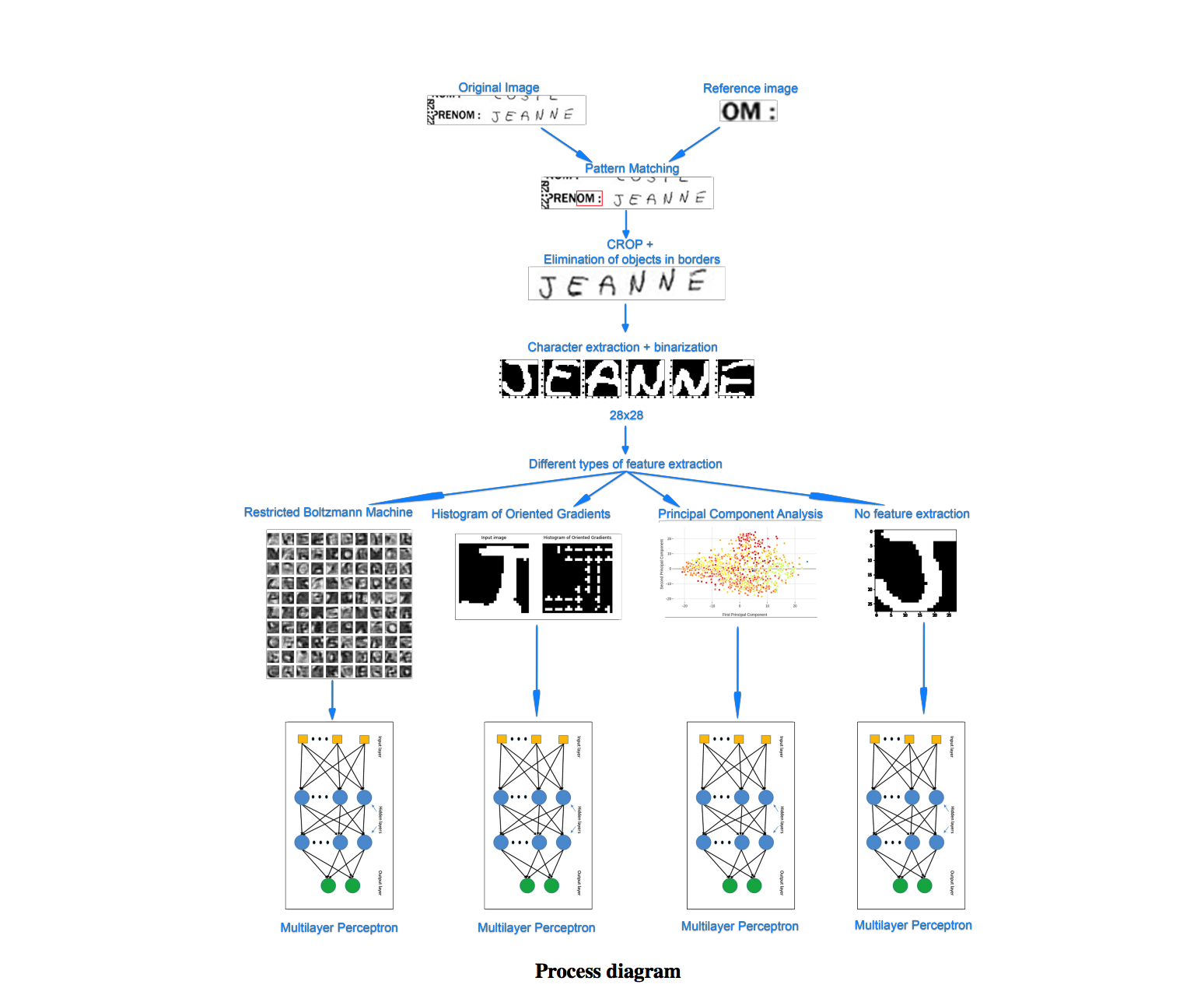
* Recognize digit
* Recognize letter
* Recognize multiple digits
* Recognize multiple letters
* Recognize handwriting



Let's explore how we can extract text from any forms / scanned pages.

**Objectives**

The goal is to find an algorithm that can extract the maximum information from a given page (jpg format). So, we can feed it to another system. (Business logic, neural network, classifier, etc.) The overall process may not be perfect. But it would be great if it can find enough information to identify the type of document and the involve identities.

* Parse any form / scanned page and extract any text data (printed text and handwriting text). So, no prior knowledge of the layout / structure of the document.
* Automatic extraction process (no human interaction. So, it can scale out)
* Somehow fast (or the ability to speed up the task with more machines or CPU)

**Challenges**

There are many challenges to overcome. But the main problem is to identify which part of the form contains text.

Some other challenges:

* Black Border Removal
* ICR (Intelligent Character Recognition): recognize and convert hand-drawn characters into text
* Scanned page (Detect edges and apply a perspective transform to obtain the top-down view of the document)
* Remove noise (blur, OTSU, adaptivethreshold with opencv)
* Shape detection and extraction
* OCR (Not a real issue since we can use : Tesseract 4 great for printed text)
* Handwriting recognition
* Minimize errors

The process of ICR is broken in to the following 7 steps:

* Character isolation
* Noise reduction
* Boundary removal
* Normalising
* Thinning
* Feature extraction
* Training or Recognition

Proces of recognition is divided into 5 steps, starting with photo of page with text.

1. Detection of page and removal of background
2. Detection and separation of words
3. Normalization of words
4. Separation and recegnition of characters (recognition of words)