OPTICAL CHARACTER RECOGNITION THROUGH MACHINE LEARNING TECHNIQUES

ABSTRACT

Digital identity is the cornerstone of a complete digital experience that customers now expect. Common tasks such as enrolling for a new service, logging into an existing service, making changes to an existing account, or performing a payment all rely on customers being able to prove who they are. Without a verified digital identity, companies open themselves up to fraud, regulatory penalties and exposure to other security risks. It is with this in mind we are on this venture to develop a web based application for verifying the identity through machine learning techniques.

In this application identity of a person is scanned through a camera or device and information is extracted from this document . The information that extracted from the id is stored and cross matched with details in our hand. Through this proofing it can ensure that applicants are who they claim to be.

Here we use Tesseract algorithm for data extraction from a scanned documents. This algorithm is able to accurately de-cypher and extract text from a variety of sources! As per it's namesake it uses an updated version of the tesseract open source OCR tool. We also automatically binarize and preprocess images using the binirazation .so tesseract has an easier time de-cyphering images. Not only are we able to extract english text, but tesseract supports over 100 other languages as well