Resumen

En la actualidad, los documentos digitales se han vuelto una parte esencial de nuestra vida cotidiana. Actualmente podemos encontrar un documento digital para casi cualquier libro o documento que necesitemos, pero una gran problemática es que muchos de estos documentos digitales son imágenes guardadas en formato PDF, lo que hace muy difícil la extracción de la información de manera digital. Debido a estas y otras problemáticas se han generado sistemas de procesamiento de imágenes que busca recuperar la información almacenada mediante el Reconocimiento Óptico de Caracteres (OCR) pero una gran limitante de este tipo de sistemas es que no puede definir un Orden de Lectura lógico. El Orden de Lectura no es más que la secuencia lógica de interpretación de la información contenida en un documento. Mediante el procesamiento de documentos en formato PDF y procesamiento digital de imágenes, en este proyecto se busca desarrollar un algoritmo capaz de identificar el Orden de Lectura de un documento que permita extraer su información de forma ordenada. Esto se lo realizará en base a Lógica Difusa, la cual se basa el Razonamiento Aproximado y en el uso de Reglas Lingüísticas. Este proyecto puede ser usado para la recuperación de información y así crear bibliotecas virtuales o aplicaciones que sirvan de ayuda a personas con discapacidad visual.

PALABRAS CLAVE:

- **LÓGICA DIFUSA**
- **RAZONAMIENTO APROXIMADO**
- **BOUNDING BOX**
- **ORDEN DE LECTURA**

Abstract

Today, digital documents have become an essential part of our daily lives. Currently we can find a digital document for almost any book or document that we need, but a great problem is that many of these digital documents are images saved in PDF format, which makes it very difficult to extract the information digitally. Due to these and other problems, image processing systems have been generated that seek to recover the information stored by means of Optical Character Recognition (OCR) but a great limitation of this type of system is that it cannot define a logical Reading Order. The Reading Order is nothing more than the logical sequence of interpretation of the information contained in a document. Through the processing of documents in PDF format and digital image processing, this project seeks to develop an algorithm capable of identifying the Reading Order of a document that allows the information to be extracted in an orderly manner. This will be done based on Fuzzy Logic, which is based on Approximate Reasoning and on the use of Linguistic Rules. This project can be used for information retrieval and thus create virtual libraries or applications that help people with visual disabilities.

KEYWORDS:

- **DIFFUSE LOGIC**
- **APPROXIMATE REASONING**
- **BOUNDING BOX**
- **READING ORDER**