Resumen

En la actualidad el paso de información científica en su mayoría se da mediante artículos que en gran parte de los casos se encuentran en formato PDF, lo cual ha hecho que crezca la popularidad de dicho formato y que hace necesario manipular este tipo de documentos, tareas como extraer texto, tablas, figuras y formulas son ineludibles para ser analizadas y procesadas. Una de las tareas que se encuentran pendientes en cuanto al trabajar con archivos de este tipo es el reconocimiento de fórmulas matemáticas. Una de las tareas más importantes en la detección y reconocimiento de fórmulas matemáticas es identificar correctamente su ubicación dentro de un documento, uno de los principales problemas en todos estos trabajos dedicados a la detección de fórmulas es validar su desempeño ya que los programas y las bases de datos con las que se puede realizar la validación no son válidas o no son de uso libre. En este proyecto se busca mediante el procesamiento de archivos pdf y procesamiento digital de imágenes crear una base de datos que contenga posición y caracteres de fórmulas matemáticas extraídas de un archivo en formato PDF. Para eso se usará como base el trabajo propuesto por (Xiaoyan Lin L. G., 2012) para que nuevos algoritmos y los ya existentes de reconocimiento posición y caracteres de fórmulas matemáticas puedan ser evaluados o probados para tener un criterio equitativo de rendimiento.

Palabras Clave:

- **DETECCIÓN DE FÓRMULAS MATEMÁTICAS.**
- ***BOUNDING BOX (BBOX.)***
- **HERRAMIENTAS DE EVALUACIÓN.**
- **PROCESAMIENTO DIGITAD DE IMÁGENES (PDI).**
- **COMPONENTES CONEXOS.**

Abstract

At present, the passage of scientific information is mostly through articles that in many cases are in PDF format, which has made the popularity of this type of format grow and makes it necessary to manipulate this type of documents, tasks such as extract text, tables, figures and formulas are inescapable to be analyzed and processed. One of the tasks that are pending in terms of working with files of this type is the mathematical formula identification. One of the most important tasks in the detection and recognition of mathematical formulas is to correctly identify their location within a document, the main problems in all these works dedicated to the detection of formulas is to validate their performance because the programs and databases that they can usea are not meaningful or are not free to use. In this project with the processing of pdf files and digital image processing, we will create a database that contains a position of formulas and characters extracted from a PDF file. For this, the work proposed by (Xiaoyan Lin L. G., 2012) be used as a basis. For the evaluation of performance a metric will be proposed for the mathematical formula identification.

Keywords:

- ***MATHEMATICAL FORMULA IDENTIFICATION.***
- ***BOUNDING BOX (BBOX.)***
- ***EVALUATION TOOLS.***
- ***DIGITAL IMAGE PROCESSING.***
- ***RELATED COMPONENTS.***