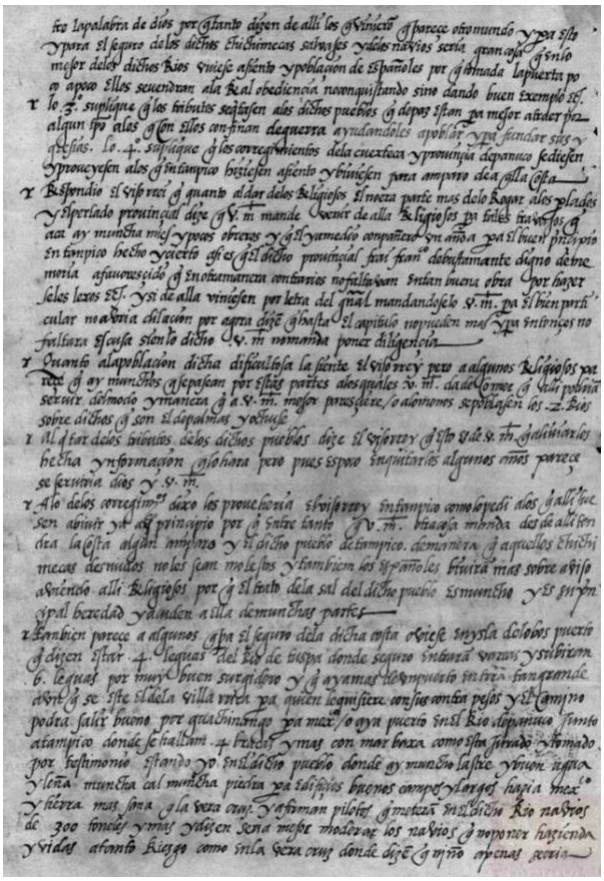
Resumen Ejecutivo del Proyecto de Obtención de Grado

CREACIÓN DE UNA BASE DE DATOS PARA EL ANÁLISIS PALEOGRÁFICO DE DOCUMENTOS EN ESPAÑOL ANTIGUO POR MEDIO DE APRENDIZAJE PROFUNDO

Dr. Iván Esteban Villalón Turrubiates, villalon@iteso.mx

1) Descripción general del proyecto propuesto

De acuerdo con la Academia Mexicana de la Lengua, el español es una lengua romance, es decir, tiene sus orígenes en el latín que abarcó gran parte de Europa, África y Asia evolucionando a través de los siglos. Es así que en España se formaron diferentes dialectos en cada región hasta que lentamente el castellano, que tuvo su cuna en Castilla, fue ganando la supremacía sobre los demás dialectos. Al conformarse España como nación, se reconoció a este idioma como español, denominación que predomina hasta estos días. La Paleografía es una ciencia que estudia escrituras históricas para identificar sus elementos y comprender su contexto, manteniendo su escritura original pero traduciéndolo a lenguaje moderno para que pueda ser comprendido de manera sencilla. Existen distintos sistemas computacionales capaces de analizar y comprender escritura antigua en el idioma inglés, sin embargo, el análisis de textos antiguos en español es un área poco explorada. Actualmente, los paleógrafos que trabajan con documentos en español utilizan sus habilidades personales basadas en sus experiencias para analizar e interpretar de manera manual los textos escritos en español antiguo, lo cual lleva tiempo y puede derivar en errores.



El proyecto desarrollará una metodología basada en el procesamiento digital de imágenes extraídas de documentos que contienen texto en español antiguo, para que puedan ser analizados e interpretados por una red neuronal convolucional la cual haya sido entrenada con los símbolos del español antiguo y, con ello, contar con un sistema automático de reconocimiento de texto y su traducción a palabras y símbolos en español moderno que puedan ser fácilmente interpretados por un paleógrafo. Este proyecto consiste de tres etapas:

1. La creación de una base de datos que será empleada para el proceso de entrenamiento de la red.
2. El proceso de pre-procesamiento de las imágenes que contienen los documentos digitales.
3. El proceso de identificación de elementos empleando la red entrenada.

La figura muestra una página de una carta de Fray Andrés de Olmos, un cura Franciscano, enviada al Emperador de España para agradecerle por su promesa de no disponer de los nativos de la región de Tampico en México, así como de las provincias de los Chichimecas donde su orden estaba realizando labores. La carta completa (de 8 páginas) es del mes de noviembre de 1556.

2) Objetivo General

En este trabajo de obtención de grado (TOG) se propone la creación de una base de datos de elementos y símbolos extraídos de documentos digitales que contienen texto en español antiguo, los cuales serán empleados para el proceso de entrenamiento para su análisis e interpretación por medio de una red neuronal convolucional, para con ello contar con un sistema automático de reconocimiento de texto y su traducción a palabras y símbolos en español moderno. Se llevará a cabo su implementación bajo diversos escenarios de prueba (imágenes sintéticas y reales) para reducir la carga computacional y ajustar los parámetros de la red para que el procesamiento a nivel software se lleve a cabo en tiempo real.

3) Entregabe esperado

Se pretende que produzca los resultados suficientes para un producto de investigación publicable, ya sea en una revista científica con factor de impacto registrado en el JCR (Thompson's Journal Citation Report), o una publicación en conferencia internacional.

4) Vinculación o colaboración

El TOG contenido en este documento forma parte de las opciones de temas de investigación a ser realizados por los estudiantes de la Maestría en Sistemas Computacionales (MSC), los cuales son de enfoque integral y que proponen un equilibrio entre la base teórica y su aplicación práctica para la solución de problemas. Abona de manera directa al proyecto de investigación titulado "*Análisis de Datos para la Extracción de Conocimiento a Partir de Modelos Multidimensionales de Percepción Remota*", el cual está registrado en el Programa Investigación (PI) atendiendo la línea de generación y aplicación del conocimiento (LGAC) "*Desarrollo de Software de Alto Desempeño*".

5) Asignaturas de la MSC relacionadas con el desarrollo del proyecto

Este TOG atiende los conocimientos obtenidos en las materias fundamentales de la MSC, los cursos de "Investigación, Desarrollo e Innovación (IDI)", así como las materias del área electiva: "Programación para Análisis de Datos", "Aprendizaje Automático (Machine Learning)", "Aprendizaje Profundo (Deep Learning)" y "Bases de Datos Avanzadas".

6) Participación en el proyecto

Este TOG pretende contar con un alumno de la Maestría en Sistemas Computacionales.

Bibliografía relacionada

[1] A. H. Alkilani and M. I. Nusir, "An Automatic Paleography Script Recognition System for the Arabic Language based on Fast Independent Component Analysis (Fast-ICA) and Support Vector Machine (SVM)," *2021 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, Amman, Jordan, 2021, pp. 49-54, doi: 10.1109/JEEIT53412.2021.9634154.

[2] S. Faigenbaum-Golovin, A. Shaus and B. Sober, "Computational Handwriting Analysis of Ancient Hebrew Inscriptions—A Survey," in *IEEE BITS the Information Theory Magazine*, vol. 2, no. 1, pp. 90-101, 1 Oct. 2022, doi: 10.1109/MBITS.2022.3197559.

[3] A. Bria *et al.*, "Deep Transfer Learning for writer identification in medieval books," *2018 Metrology for Archaeology and Cultural Heritage (MetroArchaeo)*, Cassino, Italy, 2018, pp. 455-460, doi: 10.1109/MetroArchaeo43810.2018.9089780.

[4] V. Romero, J. A. Sánchez and A. H. Toselli, "Active Learning in Handwritten Text Recognition using the Derivational Entropy," *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Niagara Falls, NY, USA, 2018, pp. 291-296, doi: 10.1109/ICFHR-2018.2018.00058.

CV del proponente

Dr. Iván Esteban Villalón Turrubiates (IEEE Student 2003, Member 2005, Senior Member 2012), obtuvo el Título como Ingeniero Mecánico y el Grado de Maestro en Ciencias en Ingeniería Eléctrica con especialidad en Procesamiento Digital de Señales, ambos por la Universidad de Guanajuato (UG) Campus Salamanca en los años 2000 y 2003, respectivamente. También obtuvo el Grado de Doctor en Ciencias en Ingeniería Eléctrica por el Centro de Investigación y de Estudios Avanzados (CINVESTAV) Unidad Guadalajara en el año 2007. Actualmente se desempeña como Profesor e Investigador Titular de tiempo completo en el Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO) en Tlaquepaque Jalisco, teniendo a su cargo labores de docencia e investigación científica, además de la Coordinación Docente de Ingeniería de Software y la Coordinación de la Maestría en Sistemas Computacionales (MSC) para el periodo 2019 a 2024, la cual tiene orientación profesionalizante y está registrada en el Sistema Nacional de Posgrados (SNP) del CONAHCYT con referencia 003869. Su trabajo de investigación está enfocado en aplicaciones del procesamiento digital de señales e imágenes a datos multiespectrales e hiperespectrales de percepción remota, a partir del cual ha publicado numerosos trabajos en revistas indexadas, conferencias internacionales de alto impacto, capítulos en libros y reportes técnicos, entre otros. Es miembro fundador y presidente del Capítulo Profesional de la Sociedad de Geociencia y Percepción Remota (Geoscience and Remote Sensing Society, GRSS) Sección Guadalajara del Instituto de Ingenieros Eléctricos y Electrónicos (Institute of Electrical and Electronics Engineers, IEEE).