UNIVERSIDAD FASTA

FACULTAD DE INGENIERÍA

Trabajo Final de Graduación

Resumen

AREX-TI es un sistema desarrollado para facilitar y ayudar a los peritos informáticos en el proceso de análisis de pericias. El mismo automatiza el análisis, el reconocimiento y la extracción de texto en imágenes, agilizando los tiempos del proceso pericial y mejorando los resultados obtenidos con respecto a las prestaciones actuales.

Título del Proyecto

AREX-TI

Autor/es: Angelucci Javier, Cassettai Pablo, Cortinez Mariano

Correo Electrónico: javierangelucci@gmail.com
pc@ufasta.edu.ar
mcortinez88@gmail.com

Director Técnico: Ing. Martín Alfredo Castellote

Director Funcional: Ing. Santiago Trigo

Año de Presentación: 2020

Abstract

AREX-TI es un sistema desarrollado para facilitar y ayudar a los peritos informáticos (y a quienes lo utilicen a posteriori) en el proceso de análisis de pericias. El mismo automatiza el análisis, el reconocimiento y la extracción de texto en imágenes, agilizando los tiempos del proceso pericial y mejorando los resultados obtenidos con respecto a las prestaciones actuales. Además, devuelve información enriquecida al usuario (tipo de imagen, fecha de la imagen, ubicación geográfica, usuario creador, dispositivo, etc) la cual es posible visualizar a través de distintos reportes; ya sea para realizar búsquedas de palabras específicas o brindar enfoques diferentes pero complementarios de la misma información, todo esto, cumpliendo con las leyes de protección de datos personales.

Palabras Claves

Informática Forense, Pericias Informáticas, Análisis de imágenes, Reconocimiento de imágenes, Extracción de texto

Introducción

1.1 Propósito

Facilitar y ayudar a los peritos informáticos en el proceso de análisis de pericias judiciales.

1.2 Problema

Se recibe un gran volumen de imágenes provenientes de una investigación judicial y se les solicita a los peritos informáticos que puedan buscar una palabra o texto específico dentro de las mismas.

El caso particular que dio origen a este proyecto fue que al cliente, perito informático, se le pidió analizar un set de imágenes (setenta mil) que era imposible realizarlo de forma manual, en busca de una palabra específica. Esto hizo tomar consciencia de la necesidad de desarrollar una solución que no solo automatice esta tarea sino que además permita extender la búsqueda a un conjunto de palabras y/o frases.

1.3 Fundamentación

Actualmente, la problemática presentada se resuelve mirando imagen por imagen, seleccionando aquellas que contenga la palabra buscada y transcribiendo a mano una a una. En grandes volúmenes de imágenes (por ejemplo, más de diez mil), realizar este proceso manual no solo introduce errores humanos (pérdida de evidencias), sino también

que conlleva a un desgaste mental; reducción en la capacidad de atención y análisis de quién/es realizan esta acción.

Por lo tanto es necesario una solución que:

- Automatice el proceso que actualmente se realiza de forma manual analizando visualmente cada una de las imágenes.
- Desarrolle un módulo orientado a segmentar, es decir, extraer las regiones donde hay texto de las capturas de imágenes de chat (esto permite orientar a las herramientas de OCR para mejorar la precisión en sus resultados).
- Favorezca el desarrollo modular, de manera que se le pueda implementar plugins en el futuro para extender la funcionalidad y/o que pueda acoplarse con otros proyectos del InfoLab como por ejemplo el proyecto de Procesamiento de Lenguaje Natural (NLP).
- Se ejecute de forma local sin la necesidad de subir información sensible a la nube. Por las leyes de protección de datos personales, sería cuestionable cargar la información de una investigación penal a un sistema alojado en servidores extranjeros (Dropbox, AWS, Google o similares).
- Permita ahorrar tiempo al generar resultados más rápidamente, algo de vital importancia en una investigación judicial.
- Guarde los resultados en una base de datos que vincule el texto encontrado con la imagen origen. Así, se incrementará el potencial del sistema en cuanto a las búsquedas (por medio de keywords, expresiones regulares y/o NLP) y posibilite su extensión.
- Que permita a los desarrolladores hacer uso de la información almacenada en la base de datos y vincularlos con otros sistemas.
- Presentación de la salida de los datos de forma amigable y detallada para el usuario.

1.4 Objetivo general

Obtener un producto multiplataforma que:

- Automatice el reconocimiento, el análisis y la extracción de texto en imágenes.
- Optimice la capacidad de búsqueda de una palabra y/o texto específico dentro de las mismas.
- Devuelva información enriquecida al usuario (Por ejemplo, fecha de la imagen, ubicación geográfica, usuario creador, dispositivo, otros metadatos.
- siempre que estén disponibles, información de la región del texto cómo el bounding box y las coordenadas del mismo, etc).

- Sea extensible a futuras funcionalidades.
- Sea fácil de integrar con otros sistemas existentes en el Info-Lab.

1.5 Objetivos específicos

1. Reconocimiento y detección automática de regiones de texto en capturas de imágenes de chat.
2. Extracción automática de texto en las capturas de imágenes de chat.
3. Indizar el texto extraído en una base de datos, para poder realizar búsquedas.
4. Realización de reportes.
5. Análisis de la aplicabilidad de (1) y (2) a otros dominios de problemas relacionados. (Imagenes en general)

1.5. Conclusiones

El proyecto no solo fue un gran desafío técnico, sino que además fue un esfuerzo de investigación para analizar y estudiar la aplicabilidad de las distintas tecnologías y herramientas existentes al contexto presentado en dicho documento.

Los objetivos planteados en un principio fueron considerablemente ambiciosos para el tiempo que teníamos para desarrollar el proyecto, pero en mayor o menor medida, todos ellos fueron logrados.

Fue un año complejo en cuanto a gestión de tiempos. Dejamos de tener la rutina y obligación de ir a cursar y ponerse a estudiar. En contrapartida tuvimos que estudiar más que antes para los finales pendientes y el desarrollo del proyecto.

Nuestra gestión de tiempo, implicaba trabajar en el proyecto 1:30 hrs todos los días sin distinción de días de estudio, fines de semana, vacaciones y demás. Esfuerzo que se incrementó a 2 hrs por día en el momento de la replanificación. A esto se le sumaba las horas laborales de cada integrante del equipo, las horas destinadas a estudio de finales/exámenes y las horas personales de cada uno.

Por lo anteriormente dicho, es destacable y por eso valoramos la solidaridad y el trabajo en equipo. En los momentos difíciles en los cuales algún integrante estaba estudiando para rendir un final y no podía dedicarle tiempo al proyecto, estaban los otros cubriéndolo.

Cuando quedaba trabado en algún problema y no encontraba solución estaba el resto de equipo colaborando para poder salir de dicha situación. Así fue cómo intercambiamos roles y liderazgo durante todo el desarrollo del proyecto y fue donde explotamos distintas habilidades de gestión de cada miembro. Las relaciones humanas también crecieron ya que no solo nos consolidamos como equipo sino que lo hicimos también como amigos, colegas y profesionales.

Todo este camino de sacrificios, aciertos, desaciertos, privaciones, alegrías y un sin fin de cualidades, nos llevó a estar al final de un recorrido que no inició con este proyecto, el inicio fue mucho antes cuando decidimos elegir qué queríamos hacer de nuestras vidas y que hoy se plasma con la finalización de un proyecto que da un cierre a todo este camino del cual estamos muy orgulloso de haber transitado.

Aclaración: Este documento es un resumen del documento de memoria de proyecto del alumno. Documentación técnica adicional del proyecto se encuentra en medios alternativos como CDs, DVs, etc.