INFORME DE INVESTIGACIÓN

NOMBRE DE PROYECTO: SISTEMA PARA LA RECUPERACIÓN DE DOCUMENTOS DE IDENTIFICACIÓN INSTITUCIONAL A PARTIR DEL RECONOCIMIENTO DE PATRONES

MODALIDAD: PARTICIPACIÓN ACTIVA EN GRUPO DE INVESTIGACIÓN

GRUPOS DE INVESTIGACIÓN:

- GRUPO DE INVESTIGACIÓN EL TELEMÁTICA Y TIC APLICADA A LA EDUCACIÓN- **TELEMATICS**.
- GRUPO DE INVESTIGACIÓN EN INFORMÁTICA, ELECTRÓNICA Y COMUNICACIONES- **INFELCOM**.

ESTUDIANTES - INVESTIGADORES PRINCIPIANTES EN LOS GRUPOS DE INVESTIGACIÓN:

- LARRY MAURICIO PORTOCARRERO LÓPEZ. CÓDIGO 201311353 http://scienti.colciencias.gov.co:8081/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001663058
- ALVARO RAMIRO HERNÁNDEZ MILLÁN. CÓDIGO 201311554 http://scienti.colciencias.gov.co:8081/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000080210

TUTORÍA:

- DIRECTOR: MIGUEL ANGEL MENDOZA MORENO
- CO- DIRECTOR: ALEXANDER CASTRO ROMERO

VINCULACIÓN DE LOS ESTUDIANTES A LOS GRUPOS DE INVESTIGACIÓN:

- TELEMATICS: formalizado el 25 de octubre de 2015 como semilleros. **Anexo 1**
- INFELCOM: agosto de 2016

ACTIVIDADES COMO SEMILLEROS DE INVESTIGACIÓN (etapa previa al trabajo de grado):

- TelemaTICs: Diseño de sistema para la ejecución de pruebas de penetración a través de drones- *PPDRONE*. Semilleros: Larry Mauricio Portocarrero López, Diego Armando Sierra Sierra, Andrea Katherine Velandia Perez. Producto: código fuente del prototipo desarrollado e informe de desarrollo, dispuesto en repositorio https://github.com/larry852/ppdron. **Anexo 2**
- TelemaTICs: Desarrollo de módulos de reconocimiento facial y habla, al igual que la generación de Speach. Semilleros: Álvaro Ramiro Hernández Millán, Sergio Esteban Piña Vargas. Producto: Ponencia titulada “Reconocimiento del Habla para la Interacción con una Plataforma de Gestión de Espacios y Elementos”, en el “IV Encuentro Nacional de Grupos y Semilleros ‘Investigación e Impacto Social’”. **Anexo 3 Anexo 4.**
- TelemaTICs: Participación en las sesiones de investigación colaborativa en TelemaTICs respecto a proyectos que se llevan a cabo.

- INFELCOM: Desarrollo fase temprana de portal https://www.loencontre.co/ para la notificación de documentos extraviados y hallados, con el objetivo de identificar de manera semiautomática su propietario, basado en una comunidad de Facebook usando APIs (Microsoft Cognitive Services e IBM Watson). Semilleros: Alvaro Ramiro Hernández Millán, Larry Mauricio Portocarrero López, Andrés Mauricio Gómez Rodriguez, Andrea Katherine Velandia Perez. **Anexo 5.**

Propuesta de Trabajo de Grado:

- Aprobada en fecha 30 de octubre de 2017.
- Periodo aprobado: 27 de octubre de 2017 al 26 de octubre de 2018.
  Anexo 6

LOGROS RESPECTO A LAS ACTIVIDADES DEFINIDAS EN LA PROPUESTA DE TRABAJO DE GRADO

Introducción

A raíz de la constante publicación de imágenes con documentos de identificación institucional de un vinculado (estudiante, profesor o trabajador) de la UPTC en los grupos de Facebook pertenecientes a la misma comunidad, con el fin de realizar la devolución del documento encontrado, y por otro lado, el creciente uso de los algoritmos de aprendizaje de máquina que se evidenciaba en el desarrollo de la asignatura Inteligencia Computacional y en la cual se desarrolló una etapa inicial de loencontre.co con el uso de APIs (Microsoft Cognitive Services e IBM Watson) para la implementación de estos algoritmos y darle una aproximación al problema de pérdida de documentos de identificación, se impulsó y motivó la realización de una propuesta de investigación basada en el estudio de las técnicas necesarias para llevar a cabo la implementación de dichas APIs usando librerías de código abierto para la clasificación como transformación de imágenes y la extracción de texto. Los objetivos planteados buscan desarrollar un estudio comparado de los algoritmos más representativos, definir un modelo para la indexación de información no estructurada de Facebook, desarrollar un sistema automático basado en el diseño planteado y determinar los grados de precisión en las implementaciones de los algoritmos.

Proceso investigativo cumplido

A continuación se detallan las actividades, su descripción y productos relacionados, que dan consistencia a cada uno de los objetivos previstos en el plan de trabajo cumplido.

Actividad	Descripción	Producto
Objetivo 1: Desarrollar un estudio comparado de aplicaciones, técnicas y algoritmos más representativos para la extracción de texto, transformación y clasificación de imágenes.		
Caracterizar los algoritmos más representativos para la transformación de imágenes	Revisión de técnicas para la transformación de imágenes que conducen a mejorar las condiciones de nitidez, definición, contraste, contornos y color.	Artículo de investigación: “Mejorando el Desempeño del Reconocimiento de Texto a partir de Pipelines de Transformación de Imágenes, generados Automáticamente” para publicación en IEEE Revista Latinoamericana. **Anexo 7** Estado: en proceso de revisión de publicación en revista IEEE America Latina. **Anexo 17**
Caracterizar los algoritmos más representativos para la clasificación de imágenes	Estudio comparado de los algoritmos de aprendizaje de máquina para la	Artículo de investigación: “Comparative Study of Machine Learning

	clasificación de imágenes usando librerías de código abierto	Supervised Techniques for Image Classification using an Institutional Identification Documents Dataset". **Anexo 8** Ponencia en IV Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI 2018) Estado: ponencia realizada el 4 de octubre de 2018. **Anexo 9** Artículo aceptado para publicación en memorias IEEE Xplore. **Anexo 10**
Caracterizar los algoritmos más representativos para la extracción de texto en imágenes	Revisión de la bibliografía sobre técnicas, algoritmos y herramientas OCR	Selección de motor libre Tesseract OCR
Objetivo 2: Definir un modelo de indexación donde se incluya el reconocimiento de patrones tanto en texto como en la imagen de un documento de identificación institucional y el tratamiento de información no estructurada.		
Plantear modelo de identificación de usuarios	Flujo de extracción, caracterización e individualización de usuarios a través de la plataforma y diferentes APIs.	Diagrama de flujo para la identificación de usuarios. **Anexo 11** Artículo de investigación: "Sistema para la Recuperación de Documentos de Identificación Institucional, a partir de Algoritmos de Procesamiento de Imágenes, Aprendizaje de Máquina y Reconocimiento de Patrones" para el II Encuentro Internacional de Investigación Universitaria ENIIU. **Anexo 12** Estado: rechazado. Pendiente para postular a revista indexada.
Caracterizar el estudio de caso, centrado en los documentos de identificación	Prototipo de sistema para la recuperación de documentos de	Poster socializado en Seminario de Actualización Facultad de Ingeniería-

institucionales de la UPTC y uno de los grupos propios	identificación institucional, a partir del reconocimiento de patrones	2017. Realizado el 4 de Octubre de 2017. **Anexo 13** **Anexo 14**
Objetivo 3: Desarrollar un sistema basado en el modelo diseñado, usando como base una red social.		
Implementar transformación en Imágenes de documentos de identificación institucionales de estudio de caso	Aplicación de diversas técnicas de procesamiento digital, de manera individual y grupal para la obtención de pipelines generados automáticamente que conducen a mejorar condiciones para etapas posteriores del proceso de reconocimiento de texto. Desarrollo de toolkit web y API para tratamiento de imágenes orientados al mejoramiento del proceso de reconocimiento de texto. Repositorio de código abierto con implementación, detalles de funcionalidades y manuales técnico y de usuario.	Toolkit web (para la generación de pipelines) Image-Optimization-Pipeline-APP: https://loencontre.co:3000/ API (para conexión con loencontre.co): Image-Optimization-Pipeline-API: https://loencontre.co:8000/ Repositorio: Código fuente-Manuales: https://github.com/larry852/image-optimization-pipeline Estado: Aceptación del registro de software. **Anexo 15**
Implementar clasificador de documentos de identificación institucionales de estudio de caso	Aplicación de SVM, K-NN, BPNN, CNN y Image Retraining/ Transfer Learning con la finalidad de seleccionar el clasificador de mejor rendimiento. El Clasificador seleccionado fue Image Retraining/ Transfer Learning. Desarrollo de Toolkit Web y API de Image Retraining/ Transfer Learning conectada a lo encontre.co	Repositorio con Dataset https://github.com/AlvaroHernandezM/ImagesLoencontre Toolkit Web para la clasificación de dos clases Image-Classification-Toolkit: https://www.loencontre.co:5000/ Repositorio Toolkit Web https://github.com/AlvaroHernandezM/Image-Classification-Toolkit Repositorio de Técnica Image Retraining/ Transfer Learning:

	Repositorios de código abierto con implementación, documentación, dataset construido.	https://github.com/AlvaroHernandezM/Image-Retraining-Classification Estado: Aceptación del registro de software. **Anexo 16**
Implementar reconocimiento de texto en documentos de identificación institucionales de estudio de caso	Implementación del motor de código abierto de reconocimiento de texto Tesseract OCR. Esta implementación es integrada como servicio sobre API de transformación.	API: Image-Optimization-Pipeline API root: https://loencontre.co:8000/ Servicio: ocr-individual Url: https://loencontre.co:8000/ocr-individual/{pipeline} Método http: GET Parametro: Id pipeline Respuesta: Texto reconocido sobre pipeline
Implementar procesamiento de información no estructurada de una red social para la identificación de usuarios	Revisión de técnicas y herramientas para la obtención de información pública sobre la red social Facebook. Selección de la técnica web scraping al dar ventajas sobre políticas de privacidad aplicadas a herramientas de acceso de información como APIs oficiales de Facebook. Desarrollo de script parametrizable en el lenguaje python, que permite la extracción de información no estructurada de la lista de miembros del grupo de la institución UPTC en la red social facebook, a través	Repositorio de código abierto con la implementación y manual técnico: Scraping-Facebook-Members: https://github.com/larry852/facebook-members

	de técnica de web scraping con el uso de la librería Selenium.	
Desarrollar un sistema automático que implemente la funcionalidad de la API en el problema identificado	Actualización de aplicación “loencontre.co” mediante la implementación de APIs de tratamiento, clasificación y reconocimiento de texto.	https://www.loencontre.co/
Objetivo 4: Determinar el grado de eficiencia del sistema construido por medio de pruebas de tiempos de recuperación y desempeño del modelo planteado.		
Comprobar el nivel de exactitud del software en el caso de estudio, usando técnicas de verificación	Toma y comparación de porcentaje de similitud sobre texto original contra texto detectado con la implementación de procesos de tratamiento y extracción de texto sobre imágenes de documentos de identificación del estudio de caso. Resultados reflejados sobre artículo “Mejorando el Desempeño del Reconocimiento de Texto a partir de Pipelines de Transformación de Imágenes, generados Automáticamente”	Artículo de investigación: “Mejorando el Desempeño del Reconocimiento de Texto a partir de Pipelines de Transformación de Imágenes, generados Automáticamente” en proceso de revisión de publicación en Revista IEEE America Latina. **Anexo 7**

Aportes al programa

Ingeniero Alexander Castro: la participación activa en grupo de investigación es a mi parecer una excelente opción para los estudiantes que desean graduarse, sin embargo esta ha sido poco utilizada debido a varios factores, por ello el esfuerzo de los estudiantes Álvaro y Larry es de destacar.

Un proceso que inició hace varios años como debe ser, desde asignaturas del programa y desde uno de los semilleros de investigación de un grupo de investigación del nuestro programa y que dejó como frutos para el programa de Ingeniería de Sistemas y Computación los productos reseñados en este informe, siendo de gran relevancia una ponencia internacional, un artículo de investigación en proceso de revisión de publicación a una revista internacional y dos registros de software aprobados, estos productos serán importantes en futuros procesos de evaluación de la calidad del programa y corresponden a la misión y visión del mismo.

También, es de resaltar la cooperación entre los grupos Telematics e Infelcom, correspondiendo a la necesidad de mejorar indicadores de cooperación entre grupos dentro

del proceso de medición de grupos de investigación por parte de Colciencias. En este punto es importante destacar el papel del director del grupo de investigación Telematics PhD Miguel Angel Mendoza Moreno, quien lideró el proyecto propiciando un buen ambiente de trabajo y valorando los aportes de todos los integrantes del grupo de trabajo.

Por último, es importante volver a felicitar tanto Larry y Álvaro por su compromiso y sentido de pertenencia con el programa de Ingeniería de Sistemas y Computación.

Miguel Angel Mendoza Moreno: el cierre de cualquier proceso investigativo es un hito para un programa académico, toda vez que las estadísticas en la actualidad dan cuenta de la ínfima porción de trabajos de grado en esta modalidad, en contraposición a la graduación por métodos alternos. En esta oportunidad los grupos de investigación INFELCOM y TelemaTICs presentan los resultados de un proceso investigativo en todo el sentido del término, donde los estudiantes Larry Portocarrero y Álvaro Hernández conocieron cada grupo, se identificaron con sus dinámicas, se constituyeron como semilleros de investigación, abanderaron proyectos de investigación formativa y llevaron sus conocimientos y avances a sus asignaturas perfilando proyectos de aplicación e integración de saberes, todo ello les dió la entereza necesaria para reclamar al interior de sus grupos de investigación el rol de investigadores principiantes con el que postularon su trabajo de grado y de esa manera, dando cobertura metódica a lo propuesto, revierten al programa académico de Ingeniería de Sistemas y Computación sus frutos constatables en productos investigativos de distintas escalas: (1) Reportes de avance de Investigación, (2) Ponencia internacional, (3) Artículos de investigación tanto en proceso de revisión de publicación como en avance de publicación, (4) Registros de software ante la DNDA avalados, (5) Un portal web que brinda un servicio a una comunidad virtual, accesible, funcional e instaurado y (6) Un repositorio de datos; desde luego, en la vertiente de los intangibles se encuentran elementos destacados como: (1) La experiencia investigativa, (2) El crecimiento que dan a sus grupos de investigación, (3) El know- how para realzar las competencias técnicas y profesionales, (4) La identificación de procesos exitosos y no exitosos tendientes a generar productos investigativos y (5) La demostración que la integración de saberes tratados en diferentes líneas de grupos de investigación no puede entregar cosas diferentes a la sinergia cuando se tiene la buena voluntad y disposición, como la expuesta por cada uno de los integrantes de este proceso (estudiantes e ingeniero Alexander Castro Romero). Sea este un momento destacable para el programa académico y sus grupos de investigación.

Valga la pena aclarar que el proceso investigativo que se presenta como resultado en este informe no se corresponde con alguna asignatura del programa académico, de modo que el procesamiento de imágenes no es tratado dentro del currículo pero sí amerita una relevancia sustantiva dentro de las competencias del Ingeniero de Sistemas y Computación; de esta manera se evidencia un complemento adecuado entre academia (al desarrollar proyectos de asignatura guiados al tema de investigación), investigación (por el proceso cumplido) y extensión (al brindar a la comunidad virtual de la UPTC un API completamente funcional que contribuye a tratar una problemática presentada).

Análisis del proceso cumplido

Conceptos emitidos por los estudiantes:
Las experiencias de socialización llevadas a cabo, mediante la modalidad de póster y ponencia, nos han permitido continuar creciendo en los desarrollos de investigación proyectados, como el paso a seguir; ya que la metodología de investigación llevada a cabo y los resultados alcanzados hasta ese momento, son compartidas y realimentadas con la comunidad científica interesada en los mismos campos de investigación. Entendiendo al proceso de investigación como una construcción de peldaños a través del tiempo que se desarrolla en una línea de investigación y se nutre de comunidad científica.

Respecto al póster socializado en el Seminario de Actualización Facultad de Ingeniería, el 4 de octubre de 2017, la dinámica de proporcionar los resultados de avances y recibir inquietudes (nuevas formas de verlo) es mayor por su misma naturaleza, ya que el tiempo es más prolongado como las veces que se expone lo aprendido, ejercicio que inconscientemente mejora nuestras capacidades para transmitir lo realizado y seguir en mejora. En cuanto a la ponencia realizada, el 4 de octubre de 2018 en el IV Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI), el ejercicio requería una mayor preparación para su socialización, ya que se contaba con 15 minutos para exponer de forma precisa y clara el proceso de investigación, y de 5 minutos para la sesión de preguntas, en las cuales se alcanzó a dar tratamiento a dos preguntas.

El producto obtenido para el modelo de identificación de usuarios, y que fue presentado al ENIIU en la modalidad de póster, no pudo llegar a ser socializado, ya que al ser presentado obtuvimos la siguiente respuesta para el mismo: "*El formato en el cual se presentó el documento es inválido siguiendo las recomendaciones de la plantilla, ya que el encabezado y otros elementos visuales poseen un desplazamiento y/o redimensión respecto a la plantilla original, por lo cual no ha sido aprobado. Recordamos que este año, debido a la gran cantidad de trabajos postulados, el control de estilo en cuanto a formato y forma se ha vuelto más rígido, y es un factor determinante para la selección de los trabajos*" Lo que nos permite reflexionar respecto a consideraciones de forma en eventos académicos o científicos, ya que por este tipo de modificaciones no es posible llevar a cabo la muestra del proceso y los resultados de investigación, que es lo que representa el esfuerzo.

Se llevó a cabo la implementación de los algoritmos y técnicas caracterizados para su implementación con librerías de código abierto, como la generación automática de pipelines, sin contar con la infraestructura adecuada, sino la necesaria ya que la implementación requiere niveles altos de computación para alcanzar su rendimiento máximo en tiempos mínimos. Sin embargo, todo el proceso de instalación, ejecución y despliegue de los Toolkit Web y APIs resultantes, está debidamente documentado para su implementación en otra máquina y los repositorios de código como de imágenes, se encuentran disponibles para aporte a la comunidad y futuras replicaciones, dándole relevancia al uso de Software Libre.

Actualmente, nosotros a parte del tiempo dedicado al desarrollo de esta investigación, también nos encontramos trabajando como desarrolladores y podemos afirmar que el desarrollo de la presente investigación nos ha generado mejores competencias en la

búsqueda y selección de información de calidad, el uso de tecnologías nuevas, el trabajo en equipo y presentación de resultados u objetivos alcanzados de forma clara.

La dinámica de trabajo no se vió afectada por contar con dos directores de grupos de investigación distintos, los horarios de trabajo fue en mutuo acuerdo de todos los integrantes, los aporte de cada uno fueron escuchados, analizados y tomados en cuenta, siempre se buscaba hacer más claros los conceptos y esto fue gracias al moldeamiento y el papel que jugaba cada integrante, haciendo muy enriquecedor el proceso. Realmente, como estudiantes que deciden investigar, nos encontramos totalmente agradecidos principalmente con los ingenieros Miguel Angel Mendoza Moreno y Alexander Castro Romero porque nos brindaron el apoyo incondicional en cada momento del proceso, siempre nos guiaron de la forma más correcta permitiéndonos crecer en cada paso realizado, sin embargo, tenemos también una agradecimiento profundo con los demás integrantes de TelemaTICs e INFELCOM, que también hicieron parte del proceso, intercambiando conocimiento con sus consejos y recomendaciones.

Anexos

Anexo 1. Certificado de vinculación al semillero de investigación TelemaTICs
Anexo 2. Resumen de diseño de sistema para la ejecución de pruebas de penetración a través de drones- PPDRONE.
Anexo 3. Reconocimiento del Habla para la Interacción con una Plataforma de Gestión de Espacios y Elementos.
Anexo 4. Certificación participación de ponencia nacional
Anexo 5. IDENTIFICACIÓN DE PERSONA EN FACEBOOK A TRAVÉS DE CARNÉ INSTITUCIONAL UPTC MEDIANTE CLASIFICADOR DE API WATSON, FACEBOOK GRAPH API, MICROSOFT COGNITIVE SERVICES (FACE Y OCR).
Anexo 6. Respuesta Escuela de Ingeniería de Sistemas y Computación para el periodo aprobado.
Anexo 7. MEJORANDO EL DESEMPEÑO DEL RECONOCIMIENTO DE TEXTO A PARTIR DE PIPELINES DE TRANSFORMACIÓN DE IMÁGENES GENERADOS AUTOMÁTICAMENTE
Anexo 8. Comparative Study of Machine Learning Supervised Techniques for Image Classification using an Institutional Identification Documents Dataset
Anexo 9. CERTIFICADO PONENCIA EN EL IV CONGRESO INTERNACIONAL DE INNOVACIÓN Y TENDENCIAS EN INGENIERÍA CONIITI-2018.
Anexo 10. Artículo aceptado para publicación en memorias IEEE Xplore.
Anexo 11. Diagrama de flujo para la identificación de usuarios.
Anexo 12. Sistema para la Recuperación de Documentos de Identificación Institucional, a partir de Algoritmos de Procesamiento de Imágenes, Aprendizaje de Máquina y Reconocimiento de Patrones
Anexo 13. Certificado Seminario de Actualización Facultad de Ingeniería- 2017. Álvaro Ramiro Hernández Millán
Anexo 14. Certificado Seminario de Actualización Facultad de Ingeniería- 2017. Larry Mauricio Portocarrero Lopez
Anexo 15. Estado aceptado del registro de Toolkit Web Image-Optimization-Pipeline
Anexo 16. Estado aceptado del registro de Toolkit Web Image-Classification-Toolkit
Anexo 17. Estado en revisión de la publicación del artículo "Mejorando el Desempeño del Reconocimiento de Texto a partir de Pipelines de Transformación de Imágenes, generados Automáticamente"