**Sistema Automatizado para la Digitalización de Formularios Agrícolas usando OCR con Machine y Deep Learning**

Victor Aponte, Cristian Lara y Lucas Rodríguez

Ingeniería en Informática

Instituto Profesional Duoc UC

**Notas del Autor**

Victor Aponte, Cristian Lara y Lucas Rodríguez, Ingeniería en Informática en el Instituto Duoc UC, Sede de Antonio Varas, Santiago de Chile.

Este trabajo fue realizado con la ayuda del Centro de información de recursos naturales (CIREN). Cuenta con la corrección de estilo de la profesora Rocio Contreras Aguila, de Ingeniería en Informática del Instituto Duoc UC.

**Resumen**

El proyecto HortiScan aborda la problemática de la digitalización de formularios agrícolas mediante un sistema automatizado basado en reconocimiento óptico de caracteres (OCR) y algoritmos de Machine Learning y Deep Learning. Este sistema está diseñado para mejorar la precisión y velocidad en el procesamiento de grandes volúmenes de datos en sectores como la agricultura, reduciendo la dependencia de métodos manuales. El propósito de este trabajo es desarrollar una aplicación móvil que capture imágenes de formularios en terreno, los procese automáticamente y los almacene en formato digital, facilitando su acceso y edición a través de una plataforma web. Para lograrlo, se utiliza un marco teórico centrado en tecnologías de procesamiento de imágenes y aprendizaje automático. La metodología empleada incluye el desarrollo modular de la aplicación móvil y web, la implementación de un OCR avanzado y la integración de APIs dedicadas para cada plataforma. Aún en la fase inicial, los resultados esperados del proyecto incluyen la optimización de los procesos de captura y digitalización de datos, proporcionando una herramienta adaptable y escalable que puede ser implementada en otros sectores con necesidades similares.

**Palabras Clave:** OCR, Digitalización, Machine Learning, Deep Learning, Agricultura

***Abstract***

*The HortiScan project addresses the problem of digitizing agricultural forms through an automated system based on Optical Character Recognition (OCR) and Machine Learning and Deep Learning algorithms. This system is designed to improve the accuracy and speed of processing large volumes of data in sectors such as agriculture, reducing reliance on manual methods. The purpose of this work is to develop a mobile application that captures images of forms in the field, processes them automatically, and stores them in digital format, facilitating access and editing through a web platform. To achieve this, the theoretical framework focuses on image processing technologies and machine learning. The methodology involves the modular development of* *the mobile and web applications, the implementation of advanced OCR, and the integration of dedicated APIs for each platform. Although the project is still in its initial phase, the expected outcomes include the optimization of data capture and digitization processes, providing a scalable and adaptable tool that can be applied in other sectors with similar needs.*

***Keywords:*** *OCR, Digitization, Machine Learning, Deep Learning, Agriculture.*

**Introducción**

El proyecto "HortiScan" se presenta como una solución innovadora para la digitalización de datos, utilizando tecnologías avanzadas de Reconocimiento Óptico de Caracteres (OCR) mejoradas con algoritmos de Machine Learning y Deep Learning. La importancia de este problema radica en la creciente necesidad de las industrias de optimizar la captura y procesamiento de grandes volúmenes de datos, particularmente en sectores donde el manejo de documentos físicos es un desafío significativo. La propuesta de HortiScan no solo responde a esta necesidad, sino que también aborda problemas de precisión y eficiencia en la digitalización, lo que es crucial para mejorar la competitividad de las organizaciones en un mundo cada vez más digitalizado.

El estudio se relaciona estrechamente con trabajos previos en el área de digitalización y automatización de procesos, especialmente aquellos que han utilizado OCR en combinación con técnicas de aprendizaje automático. Sin embargo, HortiScan se diferencia al integrar un enfoque más robusto mediante el uso de algoritmos de Deep Learning, lo que promete una mayor precisión en el reconocimiento de caracteres y una reducción significativa de errores en la digitalización de formularios. Además, el desarrollo de una plataforma móvil y una web para la captura, procesamiento y gestión de los formularios digitalizados expande el alcance de la solución, permitiendo una integración más fluida entre las operaciones de campo y la administración de datos en oficinas centrales, como en el caso del CIREN.

Las hipótesis y objetivos del estudio están fundamentados en la premisa de que la implementación de un sistema automatizado de digitalización con OCR mejorado puede incrementar significativamente la eficiencia y precisión en el manejo de documentos físicos en la industria hortícola. El objetivo principal es desarrollar un sistema de digitalización automatizada que optimice la captura, procesamiento y almacenamiento de datos, garantizando la precisión y eficiencia en su digitalización. Los objetivos secundarios incluyen la implementación de aplicaciones móviles y web para la gestión de los formularios, la integración de un OCR avanzado, y el diseño de una base de datos escalable. Estos objetivos están alineados con teorías de eficiencia operacional y automatización de procesos, y se espera que los resultados del estudio proporcionen una base sólida para futuras investigaciones en la optimización de la digitalización en otros sectores industriales.

El diseño de la investigación y las hipótesis están intrínsecamente relacionados, ya que la estructura modular del proyecto permite abordar cada componente de manera independiente, asegurando que los avances en un área no se vean comprometidos por los desafíos en otra. Esta metodología no solo facilita la integración final de todos los componentes del sistema, sino que también permite una adaptación rápida a los resultados obtenidos durante el desarrollo.

Finalmente, las implicaciones teóricas y prácticas de este estudio son significativas. Teóricamente, el proyecto contribuirá al cuerpo de conocimiento existente sobre la digitalización automatizada y el uso de OCR mejorado en la captura de datos. Prácticamente, HortiScan proporcionará una herramienta eficiente y precisa para la digitalización de formularios, con el potencial de ser adaptada y aplicada en otros contextos industriales, lo que podría transformar la manera en que se gestionan los datos en diferentes sectores. En resumen, este proyecto no solo tiene la capacidad de resolver un problema práctico existente, sino que también abre nuevas vías para la investigación y desarrollo en la automatización de procesos.

**Método**

**Enfoque General**

El proyecto "HortiScan" se centra en la creación de un sistema de digitalización automatizada de formularios hortícolas mediante el uso de tecnologías avanzadas en Reconocimiento Óptico de Caracteres (OCR) mejoradas con algoritmos de Machine Learning y Deep Learning. El enfoque metodológico adoptado sigue una estructura iterativa y modular, basada en principios de desarrollo ágil como Scrum, lo que permite la validación continua y mejora de cada componente del sistema. Este enfoque metodológico está respaldado por estudios que han demostrado la eficacia de las redes neuronales en la mejora de sistemas OCR, como se observa en el trabajo de Obaid et al. (2016). Este proyecto no solo aborda la necesidad inmediata de digitalización eficiente en la industria hortícola, sino que también tiene el potencial de optimizar procesos en diversas industrias.

Además, la implementación de técnicas de Deep Learning, como las redes neuronales convolucionales (CNN), permite una mayor adaptabilidad y precisión en la interpretación de datos complejos. Este enfoque ha mostrado resultados prometedores en investigaciones recientes, como la de Yu y Xia (2020), quienes desarrollaron un sistema de reconocimiento de caracteres altamente eficiente utilizando Deep Learning.

**Participantes y Datos**

En "HortiScan" no se involucra a participantes humanos, sino que el enfoque se centra en la captura, procesamiento y análisis de datos provenientes de formularios físicos utilizados en operaciones hortícolas dentro de la Región Metropolitana. Estos formularios contienen información crucial para la gestión y toma de decisiones en el sector, y su digitalización se lleva a cabo a través de una aplicación móvil desarrollada específicamente para este proyecto. Esta aplicación captura imágenes en campo, que luego son procesadas en las oficinas centrales del CIREN, garantizando un control riguroso sobre la calidad y seguridad de la información.

Los datos recolectados son procesados mediante un sistema OCR que integra redes neuronales profundas, permitiendo un reconocimiento preciso de caracteres, tanto manuscritos como mecanografiados. Este sistema se apoya en estudios previos que han demostrado la eficacia de las redes neuronales en el procesamiento de grandes volúmenes de datos textuales, como se describe en Memon et al. (2020). Además, se toma en cuenta la literatura reciente sobre el reconocimiento de texto en imágenes, como lo documentan Obaid et al. (2016), quienes enfatizan la importancia del preprocesamiento de imágenes para mejorar la precisión del OCR en condiciones variables.

**Procedimientos de Muestreo**

El muestreo de los datos es un componente crítico en el desarrollo de "HortiScan". Se ha diseñado un plan de muestreo que garantiza la inclusión de una amplia variedad de formularios, los cuales reflejan las diferentes condiciones que pueden encontrarse en el campo, tales como variaciones en la calidad de la escritura, condiciones de iluminación y tipos de papel. Esta diversidad es esencial para entrenar el sistema OCR y asegurar su robustez y adaptabilidad en situaciones reales.

Los formularios son seleccionados en base a criterios predefinidos que incluyen la complejidad de la escritura y la diversidad de los datos capturados, asegurando así que el conjunto de datos utilizado para entrenar las redes neuronales sea lo suficientemente diverso como para generalizar correctamente durante la fase de implementación.

**Diseño de la Investigación**

El diseño de la investigación sigue una estructura experimental dividida en varias fases: captura de datos, preprocesamiento de imágenes, extracción de características, clasificación y digitalización final de los formularios. Cada fase ha sido cuidadosamente planificada para asegurar que los resultados obtenidos sean válidos y replicables.

La captura de datos se realiza mediante una aplicación móvil diseñada para optimizar la calidad de las imágenes, asegurando una alta resolución y un adecuado contraste. Esta aplicación utiliza algoritmos avanzados de procesamiento de imágenes para asegurar que los datos capturados sean de la mejor calidad posible antes de ser procesados por el sistema OCR. Según Yu y Xia (2020), la calidad de los datos iniciales es crucial para el rendimiento del sistema en etapas posteriores.

El preprocesamiento de las imágenes incluye técnicas como filtrado de ruido, segmentación y ajuste de contraste, todas ellas esenciales para preparar los datos para la extracción de características. Obaid et al. (2016) han demostrado que estos pasos son fundamentales para mejorar la precisión del OCR, especialmente en situaciones donde la calidad del input es variable. La fase de extracción de características se basa en la aplicación de redes neuronales convolucionales (CNN), que permiten identificar patrones relevantes en los datos capturados, optimizando así el rendimiento del sistema en la etapa de clasificación.

La clasificación y digitalización se realizan utilizando modelos de Deep Learning entrenados con los datos muestreados. Estos modelos permiten una clasificación precisa de los caracteres, incluso en situaciones de escritura manuscrita compleja o cuando los formularios contienen ruido visual significativo. El uso de CNN para esta tarea ha sido validado en estudios recientes como el trabajo de Yu y Xia (2020), quienes demostraron la eficacia de estas técnicas en la digitalización de documentos manuscritos.

**Medidas y Covariables**

Para evaluar la efectividad del sistema "HortiScan", se han definido varias métricas clave. La principal medida de éxito es la precisión del OCR, que se evalúa mediante la tasa de error por carácter (CER) y la tasa de reconocimiento correcto (CRR). Estas métricas son ampliamente utilizadas en la evaluación de sistemas de reconocimiento de caracteres y permiten una comparación directa con otros estudios en el campo. La tasa de error por carácter es particularmente relevante en contextos donde la exactitud del texto digitalizado es crítica, como lo han señalado estudios anteriores sobre OCR mejorado con técnicas de Machine Learning (Memon et al., 2020; Obaid et al., 2016).

Adicionalmente, se han identificado varias covariables que podrían influir en el rendimiento del sistema, tales como la calidad de la imagen (resolución, iluminación), el estilo de escritura (variabilidad en la caligrafía) y las condiciones de captura en el campo. Estas covariables son monitoreadas y ajustadas durante el preprocesamiento de las imágenes para asegurar que no afecten negativamente el rendimiento del sistema.

**Manipulaciones Experimentales**

Las manipulaciones experimentales incluyen una serie de técnicas avanzadas de preprocesamiento de imágenes y optimización de modelos de Deep Learning. En la fase de preprocesamiento, se aplican técnicas de filtrado de ruido, ajuste de contraste y segmentación para mejorar la calidad de los datos antes de la extracción de características. Estas técnicas han sido validadas en la literatura, como se detalla en los trabajos de Bagwe et al. (2020).

El sistema OCR desarrollado utiliza redes neuronales convolucionales (CNN) que han sido optimizadas específicamente para manejar la variabilidad de los datos capturados en campo. Este enfoque ha sido validado en estudios como el de Yu y Xia (2020), quienes demostraron que las CNN pueden mejorar significativamente la precisión del OCR, incluso en entornos con condiciones de captura desafiantes. Además, se han explorado diferentes configuraciones de redes neuronales para identificar la configuración óptima que maximice la precisión del sistema, siguiendo las recomendaciones de Obaid et al. (2016).

Finalmente, se han realizado pruebas extensivas para validar el rendimiento del sistema en condiciones reales, utilizando un conjunto de datos de prueba representativo que incluye una amplia variedad de formularios. Estas pruebas permiten ajustar los parámetros del modelo y garantizar que el sistema sea capaz de generalizar correctamente a nuevos datos, tal como se describe en estudios previos sobre la validación de modelos de OCR (Memon et al., 2020; Yu y Xia, 2020).

**Resultados**

Para el desarrollo del proyecto "HortiScan", se seleccionaron cuidadosamente diversas herramientas de software y tecnologías para satisfacer los requisitos técnicos del proyecto. A continuación, se describen los lenguajes y tecnologías específicos utilizados en cada componente del sistema.

**TypeScript**: Dentro del framework Angular, tanto para la aplicación móvil desarrollada con Ionic como para la plataforma web, se utilizó TypeScript como el lenguaje principal de programación. TypeScript es un superconjunto de JavaScript que añade tipos estáticos opcionales y otras características que facilitan el desarrollo de aplicaciones a gran escala. Su integración con Angular permite una mayor robustez en el desarrollo, mejor detección de errores durante la compilación y una experiencia de desarrollo más estructurada y mantenible.

**HTML y CSS**: HTML (HyperText Markup Language) y CSS (Cascading Style Sheets) son utilizados para estructurar y diseñar las interfaces de usuario tanto en la aplicación móvil como en la plataforma web. HTML define la estructura del contenido, mientras que CSS controla la presentación visual, asegurando que las interfaces sean responsivas y accesibles desde diferentes dispositivos. Estos lenguajes permiten crear interfaces de usuario atractivas y funcionales que son críticas para la usabilidad del sistema.

**JavaScript**: Aunque TypeScript es el lenguaje principal, JavaScript también se emplea dentro de Angular e Ionic para manejar ciertos aspectos dinámicos de la interfaz de usuario y para ejecutar scripts donde se requiere un comportamiento más inmediato en el navegador. JavaScript es esencial para la manipulación del DOM y para integrar funcionalidades interactivas en tiempo real en las aplicaciones web y móviles.

**SQL (Structured Query Language)**: Para la gestión de la base de datos, se utilizó PostgreSQL, que emplea SQL como lenguaje principal para la manipulación y consulta de datos. SQL es un lenguaje estándar utilizado para interactuar con bases de datos relacionales, permitiendo la creación, modificación, consulta y administración de datos almacenados en PostgreSQL. Este lenguaje es fundamental para garantizar la integridad y eficiencia en la gestión de los formularios digitalizados y otros datos asociados al proyecto.

Estas tecnologías se integraron de manera cohesiva para desarrollar un sistema robusto y escalable que satisface las necesidades del proyecto "HortiScan". La elección de estas herramientas no solo asegura la calidad y eficiencia del desarrollo, sino que también permite una integración fluida entre los diferentes componentes del sistema.

**Discusión**

El proyecto "HortiScan" se ha concebido como una solución avanzada para la digitalización de formularios hortícolas, utilizando tecnologías de Reconocimiento Óptico de Caracteres (OCR) y algoritmos de Deep Learning. Aunque el sistema aún está en desarrollo, es posible extraer conclusiones preliminares basadas en la metodología y las tecnologías seleccionadas, apoyadas por estudios relevantes en el campo.

Significado Teórico y Práctico de los Resultados Potenciales

Desde un punto de vista teórico, el uso de redes neuronales convolucionales (CNN) y otros enfoques de Deep Learning en OCR ha sido validado por investigaciones previas. El trabajo de Memon et al. (2020) destaca la eficacia de las CNN para el reconocimiento tanto de texto impreso como manuscrito, particularmente en sistemas que requieren una alta precisión en condiciones diversas. Este enfoque teórico es consistente con el objetivo de "HortiScan" de desarrollar un sistema robusto que pueda manejar variaciones en la calidad de las imágenes y en los estilos de escritura encontrados en formularios agrícolas.

En términos prácticos, "HortiScan" tiene el potencial de optimizar significativamente el proceso de digitalización en la agricultura, donde la gestión eficiente de datos es crucial. El proyecto propone una solución que podría reducir la dependencia de los métodos manuales, acelerando el procesamiento de datos y mejorando la precisión, como lo sugieren los estudios revisados. Esto tiene implicaciones directas en la competitividad del sector agrícola, permitiendo una toma de decisiones más rápida y basada en datos precisos.

Validación y Replicabilidad de los Hallazgos Teóricos

Aunque los resultados empíricos de "HortiScan" aún no están disponibles, la solidez teórica de las técnicas propuestas está bien fundamentada en la literatura. Obaid et al. (2016) demostraron que las redes neuronales pueden ser entrenadas para reconocer caracteres manuscritos con una alta precisión, lo que sugiere que la metodología adoptada en "HortiScan" es prometedora. Además, los trabajos de Yu y Xia (2020) confirman que las redes neuronales, cuando se aplican correctamente, pueden mejorar significativamente el rendimiento de OCR en dispositivos móviles, lo que es directamente aplicable a las necesidades de "HortiScan".

Consecuencias Teóricas y Prácticas

Las implicaciones teóricas de este proyecto son significativas, ya que contribuyen al campo de la automatización de procesos mediante la integración de OCR con técnicas avanzadas de Deep Learning. El estudio de Bagwe et al. (2020) muestra que las tecnologías de Deep Learning no solo son capaces de manejar grandes volúmenes de datos, sino que también pueden adaptarse a la complejidad de reconocer caracteres manuscritos en entornos no controlados. Esto sugiere que "HortiScan" tiene el potencial de establecer un nuevo estándar en la digitalización agrícola.

Prácticamente, una vez implementado, "HortiScan" podría mejorar la precisión y la eficiencia en la gestión de datos en el sector agrícola, y también podría adaptarse para ser utilizado en otros sectores que requieren la digitalización de documentos físicos, como la medicina y la educación. El sistema está diseñado para ser escalable, lo que permite su adaptación a diferentes necesidades industriales.

Problemas y Limitaciones Identificadas

A pesar de las expectativas positivas, existen varios desafíos que deben ser abordados para asegurar el éxito del proyecto. Un desafío importante es la variabilidad en la calidad de las imágenes capturadas en el campo, que puede afectar la precisión del OCR. Esto fue destacado por Memon et al. (2020), quienes subrayaron la importancia de técnicas de preprocesamiento robustas para mejorar la calidad de las imágenes antes de su análisis.

Otra limitación es la diversidad en los estilos de escritura manuscrita, lo que presenta un desafío significativo para el reconocimiento preciso de caracteres. Este desafío es discutido en el trabajo de Bagwe et al. (2020), quienes sugirieron la necesidad de entrenar modelos con conjuntos de datos que reflejen esta diversidad para mejorar la generalización del sistema.

Implicaciones Futuras y Generalización

Mirando hacia el futuro, "HortiScan" tiene el potencial de convertirse en una herramienta clave para la digitalización de formularios en una variedad de sectores. La modularidad y escalabilidad del sistema facilitan su adaptación a diferentes contextos industriales. Esto es crucial para asegurar que el sistema pueda generalizarse y aplicarse en entornos más amplios.

En cuanto a la generalización de los hallazgos, es importante considerar que aunque las técnicas de Deep Learning han demostrado ser efectivas en una variedad de aplicaciones, la implementación exitosa de "HortiScan" dependerá de la calibración y el ajuste cuidadoso de los modelos para las condiciones específicas del entorno agrícola. Con un enfoque metodológico riguroso, "HortiScan" podría alcanzar sus objetivos y proporcionar una solución efectiva para la digitalización en la agricultura.

Comentario Final

En resumen, aunque "HortiScan" aún está en desarrollo, las bases teóricas y metodológicas del proyecto son sólidas y están bien respaldadas por la literatura existente. El uso de tecnologías avanzadas como OCR y Deep Learning ofrece un camino prometedor para la digitalización eficiente de formularios agrícolas, con potencial para expandirse a otros sectores. Los desafíos identificados proporcionan áreas claras para la mejora continua, y las implicaciones tanto teóricas como prácticas del proyecto sugieren que "HortiScan" tiene el potencial de hacer una contribución significativa a la automatización de procesos en múltiples industrias.

**Referencias**

Bagwe, S., Shah, V., Chauhan, J., Harniya, P., Tiwari, A., Gupta, V., ... & Mehendale, N. (2020). Optical Character Recognition Using Deep Learning Techniques for Printed and Handwritten Documents. Springer.

Obaid, I. M., El-Bakry, H. M., Eldosuky, M. A., & Shehab, A. I. (2016). Handwritten Text Recognition System Based on Neural Network. International Journal of Advanced Research in Computer Science & Technology, 4(1), 72-77.

Yu, W., & Chunlei, X. (2020). A New Deep Learning-Based Handwritten Character Recognition System on Mobile Computing Devices. Mobile Networks and Applications, 25(2), 402- 411. https://doi.org/10.1007/s11036-019-01243-5

Memon, J., Sami, M., Khan, R., & Uddin, M. (2020). Handwritten Optical Character Recognition: A Comprehensive Systematic Literature Review. IEEE Access, 8, 142641-142666. https://doi.org/10.1109/ACCESS.2020.3012542

Théodore Bluche. (2015). Deep Neural Networks for Large Vocabulary Handwritten Text Recognition. Computers and Society [cs.CY]. Université Paris Sud - Paris XI. ⟨NNT : 2015PA112062⟩. ⟨tel-01249405⟩