Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 11, Mes Noviembre, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 24-31
http://publicaciones.uci.cu

Tipo de artículo: Artículo original
Temática: Soluciones Informáticas
Recibido: 13/09/17 | Aceptado: 10/11/17 | Publicado: 24/11/17

Sistema OCR para la extracción de información digitalizada proveniente de máquinas de escribir

OCR system for the extraction of digitized information from typewriters

Devorat de las Mercedes Cespedes Rodríguez [1*], José Ernesto Placeres La O[1]

[1] Facultad 2, Universidad de las Ciencias Informáticas. dmrodriguez@estudiantes.uci.cu

[*] Autor para correspondencia: dmrodriguez@estudiantes.uci.cu

Resumen

La empresa XETID como parte de las empresas cubanas que se dedican al desarrollo de la informática tiene contratos con diferentes instituciones en las cuales se ha detectado la necesidad de gestionar la información presentes en los documentos generados por las mismas, entre las cuales se encuentra la búsqueda de información en pdf y en documentos digitalizados provenientes de máquinas de escribir, a estas empresas no contar con un sistema de búsqueda que permita la extracción del contenido(texto) presente en las imágenes esto evita la posibilidad de realizar búsquedas por el contenido presente en las mismas así como la aplicación de técnicas de minerías de textos para el resumen o clasificación del contenido. La realización de un sistema que permita la extracción de la información presente en los documentos permitiendo realizar tales operaciones, las cuales tienen un grado alto de complejidad debido a que dependen en gran medida de que los escáneres con que se realizaron la captura no poseen la suficiente calidad como para poder realizar el proceso de digitalización esto conlleva a que se tenga que también realizar un estudio muy profundo en el área de las técnicas de digitalización de imágenes para poder obtener los datos presentes en los documentos digitalizados.

Palabras clave: digitalización; extracción; imágenes.

Abstract

The company XETID as part of the Cuban companies that are dedicated to the development of information technology have contracts with different institutions in which the need to manage the information present in the documents generated by them has been detected, among which is the search of information in pdf and digitized documents from typewriters, these companies do not have a search system that allows the extraction of the content

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 11, Mes Noviembre, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 24-31
http://publicaciones.uci.cu

(text) present in the images, this avoids the possibility of searching for the content present in the as well as the application of text mining techniques for the summary or classification of the content. The realization of a system that allows the extraction of the information present in the documents allowing to perform such operations, which have a high degree of complexity due to the fact that they depend to a great extent on the fact that the scanners with which the capture was made do not have enough quality as to be able to carry out the process of digitalization this entails to that one also has to carry out a very deep study in the area of the digitalization techniques of images to be able to obtain the data present in the digitized documents.

Keywords: *digitization; extraction; images.*

Introducción

En la actualidad gran parte de la información se encuentra almacenada en forma de texto, ya sea en papel o en formato digital. El texto impreso presenta limitantes como un mayor costo de producción, menor velocidad de difusión y menor alcance en relación con el texto en formato digital, lo que unido al auge de las Tecnologías de la Información y las Comunicaciones (TIC), ha llevado a que muchas personas y organizaciones realicen un proceso de digitalización del contenido impreso. (Solano, 2016).

La digitalización consiste en la transformación de la información analógica, propia de la naturaleza, en información digital apta para ser tratada por un dispositivo de cómputo. Cuando este proceso se realiza sobre texto impreso el resultado puede ser una imagen o texto digital. La primera se puede obtener con un dispositivo de captura de imagen como un escáner o una cámara digital y el segundo mediante digitalización o utilizando Reconocimiento Óptico de Caracteres (OCR, por sus siglas en inglés) sobre una imagen digital (Fernández, y otros, 2007).

OCR es una tecnología que permite identificar y reconocer el texto contenido en una imagen digital, transformándola a caracteres que pueden ser reconocidos por dispositivos de cómputo (An Introduction to the Process of Optical Character Recognition, 2013).Los sistemas OCR emplean técnicas de reconocimiento de patrones, que se ocupan de los procesos de ingeniería, computación y matemáticas relacionados con objetos físicos o abstractos, con el propósito de extraer información que permita establecer propiedades entre conjuntos de dichos objetos. Un sistema automático de reconocimiento de patrones se puede dividir en tres etapas fundamentales (Hart, 1999):

- Adquisición de datos: en la que se obtiene una representación del objeto como resultado de un conjunto de mediciones.
- Extracción de características: donde se realiza un proceso interpretativo, cuyo resultado se considera como una nueva representación del objeto de la que se extrae información relevante sobre el mismo.

- Toma de decisiones: que corresponde a la clasificación propiamente dicha o proceso de identificación.

Los sistemas automáticos de reconocimiento de patrones están siendo utilizados cada vez más en la digitalización de la información con el empleo sistemas de Reconocimiento Óptico de Caracteres (OCR del inglés Optical Character Recognition). Este es proceso dirigido a la digitalización de textos, en el cual se identifica automáticamente a partir de una imagen diferentes símbolos o caracteres que pertenecen a un determinado alfabeto, para luego almacenarlos en forma de datos, permitiendo interactuar con estos mediante un programa de edición de texto o similar. (Alfonso, 2014)

El área de aplicación de las tecnologías que utilizan OCR va desde la digitalización y transformación, a texto de documentos y libros históricos que solo están disponibles en soporte de papel, el reconocimiento de documentos que tienen cierta estructura o diseño específico, tales como talonarios y facturas hasta ámbitos como el sector bancario para realizar ingresos de cheques de forma automática.En el sector médico para escanear e introducir formularios con datos de los pacientes a la base de datos, reconocimiento de matrículas de coche en parqueos y puntos de control y en los centros de gestión documental (Application of OCR systems to processing and digitization of paper documents, 2011).

En la Empresa de Tecnologías de la Información para la Defensa (XETID), perteneciente a las Fuerzas Armadas Revolucionarias (FAR), existe un sistema informático para la clasificación automática de contenidos en formato de texto, sin embargo, este sistema es incapaz de procesar documentos digitalizados como imágenes que contenga texto elaborado a partir de una máquina de escribir, los cuales se tienen archivados con información clasificada .Esta es una necesidad inmediata para este tipo de empresas pues presenta una base de datos muy grande con varios documentos importantes de este tipo a los cuales no se les puede realizar búsquedas específicas en su contenido real y su procedimiento actualmente es muy engorroso, tampoco pueden ser clasificados, ni les permiten hacer minería de texto.

A partir de la problemática antes descrita se define como **problema de investigación**: ¿cómo realizar la extracción de información de textos de imágenes generadas de la digitalización de documentos de máquinas de escribir? Enmarcado en el **objeto de estudio:** los sistemas de Reconocimiento Óptico de Caracteres (OCR),centradoen el **campo de acción:** los sistemas OCR para transformar imágenes digitales en formato de texto. Para darle solución al problema planteado, se define como **objetivo general**: desarrollar un sistema OCR para la extracción de caracteres en imágenes digitalizadas generadas con máquinas de escribir.

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 11, Mes Noviembre, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 24-31
http://publicaciones.uci.cu

Materiales y métodos o Metodología computacional

Métodos teóricos:

1.Análisis histórico-lógico: se utilizará para realizar un análisis del estado del arte de los principales sistemas de reconocimiento óptico de caracteres, relacionados con el campo de acción, así como las tendencias y tecnologías empleadas en el desarrollo de estos.

2.Analítico sintético: se empleará para analizar la estructura general de los sistemas de reconocimiento óptico de caracteres, obteniendo el entendimiento de cada una de sus partes funcionando en total.

3.Modelación: se empleó para modelar los componentes estructurales del sistema a desarrollar.

Métodos empíricos:

1.Observación: como instrumento para adquirir conocimiento sobre el proceso de reconocimiento óptico de caracteres.

2.Entrevista: Se realizan entrevistas con el objetivo de obtener información acerca del proceso de digitalización y de los principales problemas existentes.

Ambiente de desarrollo:

Como parte del ambiente de desarrollo se emplearon las siguientes herramientas y tecnologías, así como la metodología de desarrollo de software utilizada.

Metodología de desarrollo de software:

Proceso de Desarrollo de Software (Prodesoft v1.5)

El objetivo fundamental del Proceso de Desarrollo de Software de la empresa XETID es brindar una guía para el desarrollo de un producto de software que satisfaga los requisitos de un cliente con una planificación y una estimación de recursos predecibles. El ciclo de vida de un proyecto de software desarrollado a partir de este proceso de desarrollo de software consta de cinco fases: Inicio, Modelación, Construcción, Explotación Experimental y Despliegue. Al finalizar cada una de estas fases los representantes de cada uno de los roles presentes en el proyecto evalúan si se cumplieron los objetivos definidos al iniciar el proceso (XETID, 2012).

La figura 1 muestra el ciclo de vida del proceso de desarrollo de software de la XETID.

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 11, Mes Noviembre, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 24-31
http://publicaciones.uci.cu

Figura 1. Ciclo de vida del proceso de desarrollo de Software

En la fase de Inicio se logra una visión preliminar de la problemática a resolver, también se identifica el alcance preliminar del proyecto. En la Modelación se definen los procesos del dominio del problema, se estiman los principales riegos que presenta el proyecto y se especifica la forma de mitigarlos. Se define la arquitectura del software, determinando las iteraciones para los subsistemas y módulos en las que se dividirá el producto y la forma de construcción, se realiza la planificación detallada de la siguiente fase (XETID, 2012).

Ya en la Construcción se deben aclarar los requisitos restantes y completar el desarrollo del sistema sobre una base estable de la arquitectura. En esta fase todas las características, componentes y requerimientos deben ser integrados, implementados y probados en su totalidad, obteniendo una versión estable del producto comúnmente llamada versión beta. Se realiza la planificación detallada de la siguiente fase. La fase de Explotación Experimental tiene como propósito asegurar que el software está disponible para realizar las pruebas de aceptación por un grupo de usuarios. Incluye las pruebas del producto como parte de su preparación para ser entregado, y la realización de ajustes en respuesta a la retroalimentación recibida de los usuarios. En el Despliegue se realiza la generalización del producto en las entidades y órganos según lo

aprobado en el Cronograma de implantación. Durante el proceso de implantación por lo general no es necesaria la participación de los integrantes del equipo de desarrollo (XETID, 2012).

El modelo de desarrollo de software que propone describe la secuencia de actividades a llevar a cabo para obtener una solución más efectiva, se logra al combinar el modelo basado en componentes y el iterativo e incremental. El desarrollo iterativo e incremental es un enfoque en el que el ciclo de vida está compuesto por iteraciones. En cada iteración se obtiene como resultado un incremento. En el desarrollo basado en componentes se alcanza un mayor nivel de reutilización del software, aún en contextos distintos a aquellos para los que fue diseñado. Cuando existe un débil acoplamiento entre componentes, el desarrollador es libre de actualizar y/o agregar componentes según sea necesario, sin afectar otras partes del sistema. Mientras que en el desarrollo orientado a procesos la modelación de procesos de negocio permite realizar una rápida y profunda exploración del dominio del problema, con el fin de lograr comprensión por parte del equipo de desarrollo de los procesos que se realizan actualmente en la entidad y la relación que existe entre estos (XETID, 2012).

Resultados y discusión

1. Componente que podrá identificar los caracteres en formato de máquina de escribir a partir de las imágenes segmentadas.

2. Componente que permitirá realizar el procesamiento y la segmentación de caracteres en formato de máquina de escribir.

Python: Es un lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible. Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, usa tipado dinámico y es multiplataforma. Es administrado por la Python Software Foundation.

Posee una licencia de código abierto, denominada Python Software Foundation License, que es compatible con la Licencia pública general de GNU a partir de la versión 2.1.1, e incompatible en ciertas versiones anteriores (Python Software Foundation, 2013).

Pycharm: Es multiplataforma, funcionando en Windows, MacOS X y Linux. Provee análisis de código, depurador gráfico, probador de unidad integrada, integración con sistemas de control de versiones y permite el desarrollo web con Django (JetBrains, 2016).

Visual Paradigm 5.0: es una herramienta UML profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue; y no emplea una metodología en específico. El software de modelado UML ayuda a una más rápida construcción de aplicaciones de calidad. (Pressman, 2012)

Conclusiones

El resultado del estudio realizado permitió identificar las principales herramientas libres y privadas para el análisis de Reconocimiento Óptico de Caracteres(OCR).Identificar de manera general los algoritmos y métodos utilizados actualmente para este proceso. Se realizó el estudio de las principales métricas de calidad utilizadas en los algoritmos de OCR. Se realizó una selección para la propuesta de solución en función de los algoritmos y herramientas identificados para la implementación de los componentes a desarrollar en la solución. Se seleccionaron las herramientas para el diseño de los principales componentes y arquitectura de la solución.

Referencias

1. Afonso, Javier Hernández. 2014. Sistema de detección y almacenamiento de variables visualizadas por un dispositivo mediante pantalla o display a través de un sistema de visión de bajo costo. 2014.
2. Agnihotri, Ved Prakash. 2012. Offline Handwritten Devanagari Script Recognition. India: s.n., 2012.
3. An Introduction to the Process of Optical Character Recognition. Patel, Umal. 2013. 5, Mayo de 2013, International Journal of Science and Research, Vol. 2, págs. 155-158. 2319-7064.
4. Bieberstein, Norbert. 2006. Service-oriented architecture compass: business value, planning, and enterprise roadmap. 2006.
5. Chandarana, Mayank Kapadia Jagruti. 2014. International Journal of Emerging Technology and Advanced Engineering Optical Character Recognition. 2014.
6. Confusion Matrix. Confusion Matrix. [En línea] [Citado el: 01 de 02 de 2018.] http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html.

7. Dileep, Dinesh. 2012. A feature extraction technique based on character geometry for character recognition. 2012.
8. Fernández, Juan Alonso y Morera, Maria Perpinyà. 2007. Digitalización, catalogación y recuperación de información en los archivos fotográficos un estado de la cuestión. II Premi COBDC al millor treball acadèmic. Barcelona: s.n., 2007.
9. Freisleben, Julinda Gllavata y Ralph Ewerth and Bernd. 2003. A Robust Algorithm for Text Detection in Images. 2003.
10. Grandio, Xabier. 2017. Marketing4 eCommerce. [En línea] 14 de Julio de 2017. [Citado el: 12 de enero de 2018.] https://marketing4ecommerce.net/tensorflow-que-es-y-sus-aplicaciones/.
11. Hall, Craig Larman y Prentice. 2003. Modelo del dominio. 2003.
12. Hart, R. O. Duda and. 1999. Pattern Classification and Scene Analysis. New York: s.n., 1999.