JÓVENES EN LA CIENCIA
REVISTA DE DIVULGACIÓN CIENTÍFICA

SISTEMA DE TRANSCRIPCIÓN AUTOMÁTICA DE TEXTO MANUSCRITO CON TÉCNICAS DE RTM

Cesar Alexandier Meza Guevara[1] y Martha Alicia Rocha Sánchez[2]

RESUMEN

El objetivo de este proyecto fue implementar un sistema de transcripción de texto manuscrito, en donde a partir de la entrada de una imagen de texto manuscrito se obtenga un texto editable y legible para las computadoras. Para este trabajo empleamos el HTK[3] (Hidden Markov Model Toolkit), es un herramienta de código abierto basado en modelo ocultos de Markov. Para la experimentación utilizamos el corpus "*Spanish Numbers*[4]", se trata de un corpus de texto escrito a mano sobre los nombres de los números en español. Para el análisis estadístico de los resultados y garantizar que son independientes las particiones de los datos de entrenamiento y de los datos de prueba se utilizó la técnica *cross-validation*. El sistema de transcripción obtuvo un máximo del 87.79% de tasa de reconocimiento y un promedio total de 83.52%.

PALABRAS CLAVE Reconocimiento de Texto Manuscrito, Reconocimiento de Patrones, HTK, HMM, Inteligencia Artificial, Procesamiento de Lenguaje Natural.

[1] Instituto Tecnológico de León. Av. Tecnológico, S/N, Fracc. Ind. Julián de Obregón, C.P: 37290, Guanajuato, León, Teléfono (477) 710 5200.
[2] Profesor, Instituto Tecnológico de León, División de Estudios de Posgrado e Investigación, Departamento de Sistemas y Computación, Av. Tecnológico, S/N, Fracc. Ind. Julián de Obregón, C.P: 37290, Guanajuato, León, Teléfono (477) 710 5200; Fax: (477) 711 2072; marthaalicia.rocha@itleon.edu.mx
[3] http://htk.eng.cam.ac.uk/download.shtml
[4] https://www.prhlt.upv.es/page/data

Introducción

El reconocimiento de texto manuscrito (RTM), se utiliza para reconocer, interpretar, identificar y verificar textos, el RTM lo entendemos como la tarea que transforma un lenguaje representado por su forma espacial de marcas gráficas a una representación simbólica. Esta representación simbólica será en nuestro caso el código ASCII de 8 bits utilizado en los ordenadores (Toselli, 2004).

Figura 1. Ejemplo de un Texto Manuscrito

Hoy en día a pesar de los grandes avances tecnológicos y la gran movilidad que nos ofrecen diferentes dispositivos como: laptops, tablets y smartphones, es una realidad que gran parte de la información que se maneja en la actualidad sigue siendo representada en forma manuscrita sobre papel, como pueden ser: cartas, faxes, formularios de encuestas, anotaciones, etc., en estos documentos es conveniente procesarlos y transcribirlos automáticamente mediante un ordenador, ya que si lo hacemos de la manera tradicional gastaríamos mas recursos humanos y tardaríamos mas tiempo en hacerlo, con esto se hace de manera mas eficiente este proceso. (Toselli, 2004).

El RTM ha tomado una gran importancia en diversas aplicaciones, como lo son: el del reconocimiento de cantidades numéricas en cheques bancarios (Pastor i Gadea , 2007), verificación de firmas (Suárez Hernández , Herrera Luna , & Felipe Riverón , 2008), lectura y reconocimiento de códigos y direcciones postales (Pastor i Gadea , 2007), reconocimiento de autoría de documentos manuscritos (Herrera Luna , Suárez Hernández , & Felipe Riverón , 2007).

Para adquirir los datos a procesar se distinguen dos maneras de hacerlo se han tomado las definiciones de (Toselli, 2004):

On-Line: los datos se obtienen en tiempo real mientras se escribe. El sistema RTM trata con una representación espacio temporal de los datos de entrada.

Off-Line: los datos son registrados por medio de escáneres o cámaras en forma de imágenes. El sistema de RTM trata con representaciones en un espacio de luminancia.

Actualmente en Europa se están llevando proyectos como tranScriptorium (Sánchez, Mühlberger , Gatos , Schofield , Depuydt, & Davis , 2013) donde se quiere mejorar el pre procesamiento de imágenes y técnicas de HTR, el crear enfoques desde las palabras claves y aprovechar los nuevos enfoques interactivos predictivos de HTR.

Las principales justificacion para la realizacion del proyecto son:

Necesidad de digitalizar texto manuscrito.

Necesidad de generación de aplicaciones que ayuden a cualquier usuario en la transcripción automática de sus escritos.

Grandes volúmenes de información no digitalizadas puedan procesarse de manera automática.

Iniciar con la investigación de lingüística computacional.

Requerimiento de bibliotecas a nivel mundial la digitalización de libros antiguos.

Métodos y materiales

Para la experimentación se utiliza como base la herramienta HTK, y las imágenes son obtenidas del corpus “Spanish Numbers”, todo esto basado en modelos ocultos de markov.

Un modelo oculto de Márkov o en inglés *Hidden Márkov Model* (HMM) (Jelinek, 1997). A continuación se muestra el esquema o composición básica de un modelo oculto de Márkov con sus elementos característicos tales como lo son los estados ocultos, las salidas observables, las propiedades de transición y finalmente las propiedades de salida(ver Fig.2). (Carrillo Aguilar, 2007).

Un Modelo de Márkov de capa oculta es un autómata de estados finitos en el cual concurren dos procesos estocásticos. Uno de estos procesos no puede ser observado (de aquí viene el nombre de capa oculta), mientras que el otro proceso produce una secuencia de observaciones de salida. En este último proceso, asociado a cada estado del autómata, se puede emitir una observación de un conjunto de observaciones de salida siguiendo una cierta función de probabilidad. Se asume que cada estado satisface la propiedad de Márkov según la cual, para cualquier secuencia de eventos ordenados en el tiempo, la densidad de probabilidad condicional de un evento dado, depende solamente del numero de eventos anteriores.

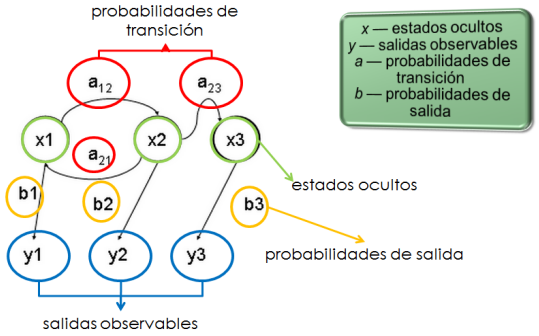

Figura 2. Esquema de un modelo oculto de Markov

La definición de un HMM (continuo) (Pastor i Gadea , 2007) (Romero Gomez, 2010) (Toselli, 2004), $\mathcal{M}$ es una máquina de estados finitos definida por la séxtupla (Q, I, F, X, a, b), donde:

- $Q=\{q_1, q_2, ..., q_p\}$ es un conjunto finito de estados.
- I es el estado inicial, es un elemento de Q: $I \in Q$.
- F es el estado final, es un elemento de Q: $F \in Q$.
- X es un espacio real d-dimensional de observaciones: $X \subseteq \mathbb{R}^d$.
- $a : (Q - \{F\}) \times (Q - \{I\}) \rightarrow [0,1]$ es una función de distribución de probabilidad de transición entre estados, tal que:

$$\sum_{q_j \in (Q-\{I\})} a(q_i, q_j) = 1 \quad \forall q_i \in (Q - \{F\})$$

- $b : (Q - \{I, F\}) \times X \rightarrow [0,1]$ es una función de densidad de probabilidad de emitir un vector $\vec{x} \in X$ en un estado $q_i \in Q$, tal que:

$$\int_{\vec{x} \in X} b(q_i, \vec{x}) d\vec{x} = 1 \quad \forall q_i \in (Q - \{I, F\})$$

En la definición de HMM dada, hay implícitos dos supuestos, Estos dos supuestos definen lo que se denomina un HMM de primer orden. (Toselli, 2004) (Pastor i Gadea , 2007) (Romero Gomez, 2010):

I. $a\,(q_i, q_j) = P\,(z_{t+1} = q_j \,|\, z_t = q_i)$ establece que la probabilidad de una cadena de Márkov en un particular estado q_j en $t+1$ depende solo del estado q_i de la cadena de Márkov en el tiempo t, y no depende de los estados visitados anteriormente en tiempos menores que t. Es decir:

$$P(z_{t+1}|z_1 \dots z_t) = P(z_{t+1}|z_t)$$

II. $b\ (q_i, \vec{x}\ ) =$ $p(x_t|x_1 \ldots x_t, z_1 \ldots z_t) = p(x_t|z_t)$ $P\ (x_t = \vec{x}\ |\ z_t = q_i\ )$ establece que la probabilidad de que $\vec{x}$ sea emitida en el tiempo t depende solo del estado q_i en el tiempo t, y no depende ni de los vectores emitidos y ni de los estados visitados anteriormente en tiempos menores que t. Es decir:

HTK es la herramienta que se empleó para el entrenamiento de los Modelos Ocultos de Márkov (Young, et al., 2002), esta herramienta a su vez nos sirve para el reconocimiento posterior de frases. Esta herramienta fue desarrollada por “Speech Vision and Robotics Group” del CUED “Cambridge University Engineering Department”, consiste en un grupo de programas, módulos y librerías escritas en código de lenguaje C.

Spanish Numbers se trata de un corpus de texto escrito a mano sobre los nombres de los números en español, este corpus fue elaborado por el "Instituto Tecnológico de Informática". De la universidad politécnica de Valencia. El corpus contiene cerca de 522 imágenes con frases de texto escrito a mano. En (Toselli, 2004) se muestra información mas especifica sobre el corpus. A continuación se muestran algunas imágenes del corpus.

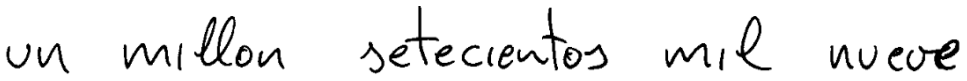

Figura 3. Imagen 010002 correspondiente al Corpus Spanish Numbers

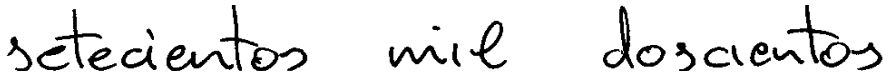

Figura 4. Imagen 010003 correspondiente al Corpus Spanish Numbers

Los experimentos de RES (Reconocimiento de Escritura.) comprenden tres fases: la de entrenamiento, la de reconocimiento y la de evaluación.

Resultados

En esta sección se presentan los resultados obtenidos durante la experimentación con la herramienta HTK, cabe destacar que el entrenamiento del modelo se hizo en base a la prueba conocida como *K-Fold*.

La validación cruzada *K-Fold* es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Lo más común es utilizar la validación cruzada de 10 iteraciones (10-fold *cross-validation*). En la tabla 1 se muestra como quedo la partición de los datos para entrenamiento y prueba.

	Entrenamiento	470 imágenes
	Prueba	52 imágenes
	Total	522 imágenes

Tabla 1. Partición de corpus para entrenamiento y prueba por k

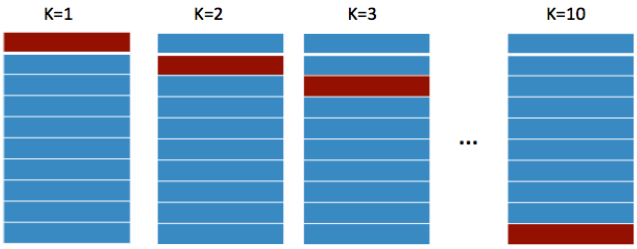

Figura 5. Ejemplo de partición de datos en un K-Fold

Resultado del entrenamiento con K=1

```
========================== HTK Results Analysis ==========================

  Date: Wed Jun 18 10:33:25 2014

  Ref : numeros-test.mlf

  Rec : res32_1.mlf

------------------------------ Overall Results --------------------------------

SENT: %Correct=28.85 [H=15, S=37, N=52]

WORD: %Corr=87.68, Acc=85.22 [H=356, D=0, S=50, I=10, N=406]
```

Tabla 2. Resultados del entrenamiento para K=1

Como se puede observar en la tabla 2 es el resultado de ejecutar el comando HResults de HTK, en la primera sección de la información cuenta con la fecha en que se ejecuto el comando, a su vez nos muestra el fichero de referencias que en este caso es el números-test.mlf, también nos muestra las hipótesis reconocidas que son almacenadas en el fichero res32_1.mlf. Ya en la parte de resultados generales (Overall Results) se muestra el significado de las letras que puede ser encontrado en (Young, et al., 2002).

En este caso en el entrenamiento para la primera iteración se puede observar que la tasa de aciertos Acc=85.22, entonces tendríamos que la tasa de error en este caso corresponde a un WER=14.78 % de reconocimiento de palabra, dado que la tasa de error se define como:

$$WER = 1 - Acc = \left(\frac{D + S + I}{N}\right).100$$

K=i	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Acc	85.22%	84.69%	84.88%	86.86%	87.79%	75.62%	83.87%	83.57%	79.49%	83.18%

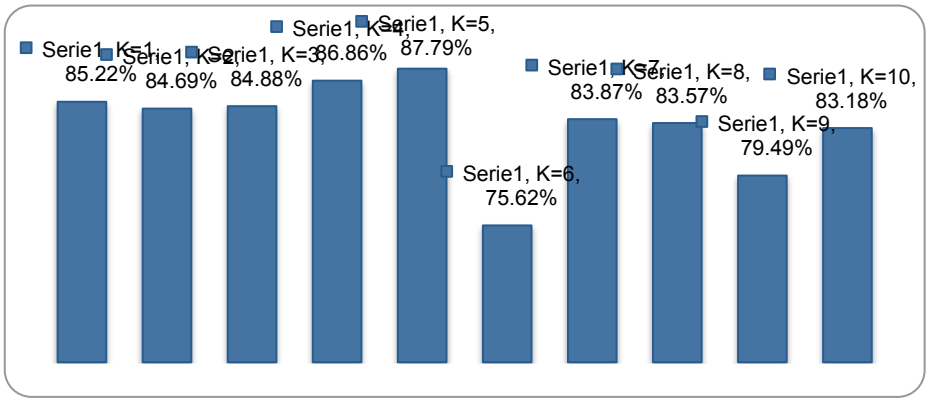

Figura 6. Resultados de la prueba k-fold

Conclusiones

Los resultados de la experimentación muestran resultados adecuados: El sistema de transcripción obtuvo un máximo del 87.79% de tasa de reconocimiento y un promedio total de 83.52%. Como trabajo futuro se requiere hacer la experimentación con diferentes corpus y desarrollar un propio extractor de características , se espera también obtener un prototipo experimental de un sistema de transcripción de texto manuscrito *online.*

REFERENCIAS

Carrillo Aguilar, R. (2007). Diseño y manipulación de modelos ocultos de Márkov, utilizando herramientas HTK. Una tutoría. *Ingeniare. Revista Chilena de Ingeniería , 15* (1), 18-26.

Herrera Luna , E. C., Suárez Hernández , D., & Felipe Riverón , E. M. (2007). *Reconocimiento de la Autoría de Documentos Manuscritos.*

Jelinek, F. (1997). *Statistical Methods for Speech Recognition.* E.U.A: MIT Press.

Pastor i Gadea , M. (2007). *Aportaciones al Reconocimiento Automático de Texto Manuscrito.*

Romero Gomez, V. (2010). Multimodal Interactive Transcription of Handwritten Text Images.

Sánchez, J. A., Mühlberger , G., Gatos , B., Schofield , P., Depuydt, K., & Davis , R. (Septiembre de 2013). tranScriptorium: A European Project on Handwritten Text Recognition. *DocEng '13 Proceedings of the 2013 ACM symposium on Document engineering* , 227-228.

Suárez Hernández , D., Herrera Luna , E. C., & Felipe Riverón , E. M. (2008). *La Firma como un Método Biométrico de Identificación .*

Toselli, A. H. (2004). *Reconocimiento de Texto Manuscrito Continuo.* Valencia, España: Universidad Politécnica de Valencia.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., et al. (2002). *The HTK Book (for the HTK Version 3.2).* Cambridge, Inglaterra: Cambridge: Entropic Cambridge Research Laboratory.