Capítulo 2.

Investigaciones Actuales en Reconocimiento de Patrones aplicado a texto

El término Reconocimiento de Patrones abarca una vasta área de procesamiento de información, desde reconocimiento de voz o de escritura cursiva, hasta detección de fallas en maquinaria o diagnóstico de enfremedades. Todos estos problemas los resolvemos los humanos de manera tan natural, que nos puede ser difícil comprender qué tan compleja es esta tarea para las computadoras.

La manera más efectiva para efectuar el trabajo de reconocer las diferentes señales de entrada ha sido la del Reconocimiento Estadístico de Patrones [Bishop 1995].

Las diferentes técnicas actuales aplicadas dentro del área de Reconocimiento Óptico de Caracteres (OCR, por sus siglas en inglés), tienen cada una sus fortalezas y debilidades. La tendencia en la investigación y en el uso de estas herramientas ha pasado de tratar de reconocer texto con tipografía de máquina a ser un problema de reconocimiento de letras con estilo cursivo (letras unidas por trazos) y texto en documentos degradados. En este contexto se sitúan las imágenes digitalizadas de las tarjetas de Hu, las cuales están degradadas por su antigüedad. Su entendimiento es difícil por el uso de fotocopias y recortes pegados en ellas, por su diferente tipografía manual y de máquina.

Dentro del área del reconocimiento de imágenes textuales existen distintos campos de investigación actual, de los cuales revisaremos 5 en este capítulo: restauración y mejoramiento de documentos degradados para optimizar el reconocimiento, el aislamiento de partes a reconocer, el reconocimiento de caracteres aislados, el reconocimiento de símbolos sin usar aislamiento, y la validación del desempeño del software de reconocimiento de texto.

2.1 Restauración y mejoramiento de documentos degradados para optimizar el reconocimiento.

El reconocimiento de texto es mejor cuando éste aparece sobre un fondo limpio y no hay otros factores que confundan al reconocedor electrónico. Sin embargo, el texto a menudo es impreso sobre fondos texturizados u obscuros. Las técnicas de mejoramiento de documentos separan los elementos textuales en una imagen, deshechan aquello que no es texto y mejoran la calidad del mismo.

La investigación llevada acabo por Wu et al. [1997] separa el texto de los demás elementos mediante heurísticas tomando en cuenta la similitud en la altura del texto, el espaciamiento entre caracteres y alíneamiento de los mismos. Este tipo de mejoramiento reconoce las partes textuales de una imagen y las aisla de las otras, pero no reconoce qué es lo que el texto significa.

Además del mejoramiento existen técnicas y teorías acerca de cómo contrarrestar los efectos físicos que influyen en la degradación de un documento. Una de las teorías de restauración de documentos es el modelado de la degradación de escritos, la cual modela mediante ecuaciones qué ocurre en el proceso de degradación de estos. Los modelos de degradación de documentos permiten introducir error (ruido) en documentos simulando el del fenómeno real, además, permiten contrarrestar los efectos de esta degradación sintética, e inclusive, restaurar los documentos afectados por estos fenómenos.

La disertación doctoral de Kanungo [1996] presenta dos modelos de degradación de documentos: (1) La degradación local de pixeles, introducida al imprimir, fotocopiar o digitalizar un documento. Esta degradación introduce defectos en los símbolos, pues al pasar de un medio a otro, partes de ellos se pierden en el proceso. (2) Un modelo físico que representa las distorsiones causadas por la perspectiva e iluminación que ocurre al fotocopiar o digitalizar un libro grueso. Este trabajo modela el fenómeno de oscurecimiento del borde del libro, y logra contrarrestar el efecto para cierta muestra de documentos que lo presentan.

Ambos modelos permiten simular en documentos reales varios niveles de degradación y restaurar los efectos de la degradación causada en ellos.

2.2 El aislamiento de partes a reconocer.

De acuerdo con Gonzales y Woods [1992], esta técnica es también llamada segmentación y consiste en subdividir una imagen en sus partes u objetos constituyentes. Esto significa que la subdivisión deberá detenerse cuando los objetos de interés hayan sido aislados. Asi, si se trata de reconocer palabras completas, la unidad mínima de aislamiento es una palabra o un signo de puntuación.

Para el caso de caracteres se puede segmentar top-down o bottom-up [Chen & Haralick 1996].

Si se procede al estilo top-down, se subdivide en áreas de texto, párrafos, líneas, palabras y caracteres, secuencialmente. Usando bottom-up se procede en forma inversa, iniciando al segmentar caracteres usando por lo general algoritmos de inundación. Al segmentar, se rodea el elemento aislado usando cajas llamadas *bounding boxes* o cajas de frontera.

Para separar algo de una imagen, es necesario primeramente definir qué es lo que lo diferencia de lo demás. En el caso de la segmentación de texto bottom-up la diferencia generalmente consiste en la variación de niveles de gris que lo representa en la imagen.

Establecido el umbral entre el texto y su contraparte, se procede a encontrar las fronteras que rodean al elemento a segmentar. Esto se logra, generalmente, al obtener un perfil de la imagen en niveles de gris y aplicarle una primera y segunda derivadas.

La primera derivada sirve para detectar los bordes y la segunda para determinar si los pixeles entre los bordes pertenecen o no al símbolo [Gonzales & Woods, 1992].

En el caso de que la diferencia de niveles de gris sea poca y suponiendo que se sabe qué niveles de gris representan texto se puede limpiar la imagen usando *thresholding*, que consiste en elevar el umbral de los niveles de gris para deshechar aquellos que no representan algo significativo.

Después de detectar los bordes se procede a determinar similitud entre los pixeles localmente para saber si pertenecen a un conjunto y así aislarlos de los demás.

Chen y Haralick proponen modelos estadísticos para adaptarse a las diferentes distribuciones de elementos de texto en documentos y así poder efectuar segmentación de una manera más sofisticada y eficiente.

2.3 El reconocimiento de caracteres aislados.

Esta técnica es la que constituye básicamente al Reconocimiento Óptico de Caracteres (OCR), y consiste en reconocer a los caracteres como símbolos aislados en una página.

El reconocimiento estadístico de patrones se usa generalmente para este propósito. Cada caracter se aisla y se representa en una matriz de pixeles, cada uno con un valor asociado de nivel de gris. Estos valores se agrupan a menudo en un vector $x=(x_1,...x_d)T$ donde *d* es el número total de tales variables y *T* determina la traspuesta. Este vector contiene las características básicas del caracter [Bishop 1995].

La meta del problema de clasificación es desarrollar un algoritmo que dado este vector de valores o características se asigne a sólamente una clase del conjunto de clases C_k. Las clases, en el dominio de esta técnica, son los caracteres ASCII. Suponemos además que se nos provee de un gran número de ejemplos de diferentes formas de escribir cada caracter. A esta colección se le puede denominar conjunto de datos (data set), en la estadística se le denomina muestra.

Es necesario entonces determinar qué características diferencian a unos caracteres de otros, pues varios vectores *x* podrían ser asignados a la misma clase, sin ser esto correcto. Es posible aumentar el número de características, pero se corre el riesgo de que así sea más difícil clasificar, pues algunas clases comparten características y la confusión aumentaría [Bishop 1995].

En contraste con la problemática que representa reconocer caracteres aislados, Gomez y Oldham [1993], definen los problemas que se presentan al tratar de reconocer texto manuscrito. Mencionan que en el tipo de letra cursiva existe una inmensa variedad de estilos, inclinaciones y tamaños, además de que pueden estar acompañados de adornos adicionales. Factores como la nacionalidad del escritor, su nivel social, educación y edad contribuyen a la poca uniformidad entre caracteres. Incluso una sola persona presenta variaciones debido al cansancio, características del medio ambiente, estado de ánimo. El principal problema derivado de esto en la implementación es el diseño del vector de características comunes entre los caracteres de la misma clase, pues la heterogeneidad de las entradas impide la definición de estas características.

2.3.1 Redes Neuronales Artificiales

Se han usado diferentes formas de abordar el problema de reconocimiento. La más popular es la simulación del funcionamiento del cerebro dentro de la computadora. Es así como surgieron las Redes Reuronales Artificiales (RNA).

Las RNA son modelos matemáticos inspirados en sistemas biológicos, adaptados y simulados en computadoras convencionales [Wasserman 1989].

Las RNA se conocen con diferentes nombres, entre los que se encuentran: modelos conexionistas, procesamiento distribuido en paralelo, sistemas neuronales artificiales, y sistemas neuromórficos.

2.6 El reconocimiento de símbolos sin usar aislamiento.

En los sistemas sin segmentación, la base del reconocimiento es la extracción de primitivas o partes fundamentales de cada símbolo [Al-Badr & Haralick 1994]. El sistema comienza por detectar un conjunto de primitivas (partes de los símbolos) y encontrar el agrupamiento óptimo de ellas en símbolos. La ventaja de este sistema es que no es necesario segmentar la imagen para obtener las primitivas. El sistema, entonces, define un símbolo como el conjunto de sus primitivas en una cierta configuración espacial. Un símbolo es reconocido en cualquier área local si tiene suficiente número de primitivas correctas, en el lugar relativo correcto. El reconocimiento libre de aislamiento surgió de la necesidad de reconocer texto a partir

de tipos de escritura que es no segmentada por naturaleza, como algunos tipos de escritura oriental.

2.7 La validación del desempeño del software de reconocimiento de texto.

En los últimos años se ha dado cierto ambiente de incertidumbre pues no hay un parámetro fijo para evaluar la eficacia de un reconocedor de caracteres con respecto a los demás. Las compañías comercializadoras de estas soluciones y los investigadores reportaban eficacia en prcentajes de 98% o 99%, pero todos ellos eran relativos a la muestra de documentos con los que se probaba, por ello se sentaron modelos para probar algoritmos reconocedores de caracteres [Kanungo 1996]. En ellos se dan modelos de verdad llamados *groundtruth*, que proveen la información real de las características del texto en un documento, como el aislamiento correcto de bloques de texto, palabras, la forma, tamaño y ubicación exacta de los símbolos en un documento. Bajo estos parámetros se puede evaluar el desempeño de las diferentes soluciones de reconocimiento.

Estas investigaciones contribuyen al continuo mejoramiento del reconocimiento de caracteres en diferentes tipos de software, ya sea comercial ó producto de investigación, la comprensión de éstos temas no sirve para efectuar una revisión de las soluciones existentes actualmente, en el siguiente capítulo revisaremos algunas de éstas soluciones.

Dircio Palacios Macedo, R. 1998. Reconocimiento y Consulta de Imágenes Textuales en Bibliotecas Digitales. Tesis Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas-Puebla. Diciembre.