INSTITUTO SUPERIOR POLITECNICO JOSE ANTONIO ECHEVERRIA

SEDE MUNICIPAL PLAZA DE LA REVOLUCION

TÃ-tulo: Reconocimiento de Caracteres Manuscritos

Grupo: 2

Curso: 2008 − 2009

Fecha de RealizaciÃ³n: 24/05/2009

RESUMEN

A travÃ©s de los aÃ±os se ha vuelto imprescindible la conservaciÃ³n de documentos y al estos hacerse viejos el papel pierde calidad, entre otros accidentes que puede sucederle. Debido a esto la ciencia ha trabajado arduamente en la soluciÃ³n de este problema inventando asÃ- diversos dispositivos electrÃ³nicos que realizan internamente mÃ©todos totalmente matemÃ¡ticos que digitalizan todo aquello que se procese, ejemplo de esto tenemos el escÃ¡ner, las cÃ¡maras digitales, entre otros.

Todo este proceso de digitalizaciÃ³n se conoce como el Reconocimiento de Caracteres Manuscritos que es el conjunto de tÃ©cnicas informÃ¡ticas con el objetivo de reconstituir los caracteres de un documento a partir de su propia imagen. En la actualidad esta disciplina cientÃ-fica no sÃ³lo engloba la reconstrucciÃ³n de caracteres, sino la estructuraciÃ³n de los documentos, y otras tÃ©cnicas de digitalizaciÃ³n.

Los sistemas de reconocimiento de caracteres comprenden diversas etapas como son: la adquicisiÃ³n de la imagen, el pretratamiento, la segmentaciÃ³n, el reconocimiento de caracteres, el reconocimiento de fuentes, el proceso de vectorizaciÃ³n, el reconocimiento de grÃ¡ficos, el reconocimiento estructural y la clasificaciÃ³n de documentos.

El reconocimiento de caracteres sigue siendo un problema complejo que tropieza con dificultades aÃºn no resueltas y que son actualmente un objeto de numerosas investigaciones.

SUMMARY

Over the years it has become essential to the preservation of these documents and to take the old paper loses quality, among other accidents that can happen. Because of this science has worked hard on solving this problem by inventing and various electronic devices that perform all mathematical methods internally digitized everything that is processed, we have the example of this scanner, digital cameras, among others.

This whole digitization process known as character recognition Manuscripts is the set of techniques with the aim of reconstituting the character of a document from his own image. At present this field of science not only includes the reconstruction of characters, but the structuring of documents and other digital technologies.

The character recognition system comprising different stages such as: the image adquicisiÃ³n, pretreatment, segmentation, character recognition, recognition of sources, the vectorization process, the recognition of graphs, structural recognition and classification documents.

The character recognition is still complex problems that is facing difficulties not yet resolved and are currently a subject of numerous investigations.

INDICE

INTRODUCCION

El problema de la mayorÃ-a de las personas e instituciones, es la conservaciÃ³n de los documentos en papel, especialmente cuando estos tienen un siglo de antigÃ¼edad. Es por esto que se ha intentado conservar tanto el documento, como la informaciÃ³n que este contiene. Un avance en la tecnologÃ-a a finales del siglo XX, ha dado origen al desarrollo de las bibliotecas digitales, las cuales ponen a disposiciÃ³n de un mayor nÃºmero de personas la informaciÃ³n. Por medio de un escÃ¡ner se puede digitalizar libros, revistas, documentos, cartas, telegramas, entre otros; pero falta un paso muy importante, interpretar el contenido e informaciÃ³n de estas imÃ¡genes digitales.

Existen en el mercado una variedad de software de reconocimiento de caracteres, sin embargo todos poseen un margen de error muy alto cuando la entrada es un documento de caracteres manuscritos. Es por esto que se da comienzo a este trabajo.

Una buena segmentaciÃ³n y reconocimiento del caracter, depende de la calidad del documento original y de la calidad de la imagen digitalizada. El principal problema en el reconocimiento de caracteres manuscritos es lo ilegible que son estos caracteres aÃºn para el ser humano.

Durante las Ãºltimas dos dÃ©cadas se han propuesto una gran cantidad de sistemas para el Reconocimiento de Caracteres Manuscritos. Sin embargo aun existen varias limitaciones relativas a su funcionamiento y porcentaje de reconocimiento sobre todo cuando los caracteres manuscritos son de tipo cursivo

El objetivo principal que se persigue con este trabajo es conocer en que consiste el mÃ©todo de Reconocimiento de Caracteres Manuscritos, y aunque este se lleva a cabo en la actualidad con distintos mÃ©todos que han ido surgiendo al pasar de los aÃ±os, aquÃ- se describe en que consiste el Reconocimiento de Caracteres, se explica el mÃ©todo principal y luego de varios de los mÃ©todos investigados se selecciona uno de ellos que en este caso es El Sistema GeomÃ©trico de Reconocimiento Ã“ptico de Caracteres Manuscritos y se desarrolla, del cual podemos decir que estÃ¡ basado en el uso de un dispositivo de exploraciÃ³n Ã³ptica que puede reconocer la letra impresa.

En la actualidad el Reconocimiento de Caracteres Manuscritos no sÃ³lo encierra la reconstrucciÃ³n de caracteres, tambiÃ©n su estructuraciÃ³n. Esta tÃ©cnica ccomenzÃ³ aplicÃ¡ndose en documentos para los cuales ninguna forma electrÃ³nica estaba disponible y a medida que la tecnologÃ-a ha evolucionado, sus aplicaciones han ido en aumento como por ejemplo las cÃ¡maras digitales cuando procesan esa foto que tomamos, las computadoras de bolsillo que vienen con pantallas portÃ¡tiles al igual que las nuevas tecnologÃ-as de telÃ©fonos mÃ³vil, Ipod, etc. El reconocimiento de caracteres sigue siendo un problema

complejo que tropieza con dificultades aÃºn no resueltas y que son actualmente un objeto de numerosas investigaciones.

DESARROLLO

IntroducciÃ³n al reconocimiento de caracteres manuscritos:

DefiniciÃ³n: Es el conjunto de tÃ©cnicas informÃ¡ticas cuyo objetivo es reconstituir los caracteres de un documento a partir de su propia imagen. En la actualidad esta disciplina cientÃ-fica no sÃ³lo engloba la reconstrucciÃ³n de caracteres, sino la estructuraciÃ³n de los documentos (tÃ-tulos, subtÃ-tulos, bloques de texto, etc...)

ComenzÃ³ aplicÃ¡ndose en documentos para los cuales ninguna forma electrÃ³nica estaba disponible. A medida que evoluciona la tecnologÃ-a, sus aplicaciones han ido en aumento. Los resultados obtenidos hasta ahora distan mucho de ser perfectos. El reconocimiento de caracteres sigue siendo un problema complejo que tropieza con dificultades aÃºn no resueltas y que son actualmente un objeto de numerosas investigaciones.

Varios factores son la causa de estas dificultades:

- Ausencia de un objetivo universal. Los resultados dependen mucho de la aplicaciÃ³n.
- Son tÃ©cnicas por lo general costosas.
- Muchas son las causas que pueden provocar que los resultados no sean los correctos.

Por ejemplo:

− resoluciÃ³n insuficiente de la imagen.

− introducciÃ³n Ã³ptica de mala calidad.

− documento deteriorado.

En general, los sistemas de reconocimiento de documentos y, por lo tanto, de caracteres comprende las siguientes etapas:

1− AdquisiciÃ³n de la imagen: mediante escÃ¡neres y cÃ¡maras.

2− Pretratamiento (BinarizaciÃ³n, Filtrado, RectificaciÃ³n).Â´

− BinarizaciÃ³n: es el proceso que toma una imagen en escala de grises y se define un umbral: los pixeles con un valor por debajo del umbral se vuelven negros, y el resto se hacen blancos. Esta es una operaciÃ³n Ãºtil en tanto que la mayorÃ-a de los problemas buscan distinguir dos objetos diferentes en la imagen: el objeto de interÃ©s y el fondo. El valor del umbral debe ser escogido de acuerdo a un criterio particular no sÃ³lo para cada problema sino para cada imagen, debido a las variaciones en la tinciÃ³n o en la iluminaciÃ³n de la muestra; en este aspecto es bastante Ãºtil la informaciÃ³n que da el histograma de la imagen. Los mÃ©todos de binarizaciÃ³n pueden dividirse en seis categorÃ-as diferentes como son:

MÃ©todos basados en la forma del histograma; sus mÃ¡ximos y mÃ-nimos, la

curvatura, etc.

MÃ©todos basados en clustering; se busca clasificar los distintos niveles de

intensidad en dos grupos diferentes.

MÃ©todos basados en entropÃ-a*(Ver Glosario, PÃ¡g. 16). Generalmente son mÃ©todos que usan la entropÃ-a del frente y el fondo de la imagen, o la entropÃ-a cruzada entre la imagen original y la binarizada.

MÃ©todos basados en atributos de objetos; se apoyan en medidas como la

coincidencia de bordes, similaridad de formas, etc.

MÃ©todos espaciales, basados en distribuciones de probabilidad o en la correlaciÃ³n entre los pixeles de la imagen.

MÃ©todos locales, que adaptan el umbral de acuerdo a las caracterÃ-sticas

de la regiÃ³n analizada.

− Filtrado: en el Ã¡rea de procesamiento de imÃ¡genes hay una gran cantidad de operaciones o filtros con efectos muy especÃ-ficos sobre la imagen aplicada; actualmente son la parte mecÃ¡nica de los trabajos en el Ã¡rea, pero no por ello menos importante. Son la soluciÃ³n mÃ¡s directa para retocar una imagen antes de intentar cualquier proceso de segmentaciÃ³n.

− RectificaciÃ³n: es el mÃ©todo que se encarga de rectificar cada uno de estos mÃ©todos antes descritos.

3− SegmentaciÃ³n: Mediante la segmentaciÃ³n se divide la imagen en las partes u objetos que la forman. El nivel al que se realiza esta subdivisiÃ³n depende de la aplicaciÃ³n en particular, es decir, la segmentaciÃ³n terminarÃ¡ cuando se hayan detectado todos los objetos de interÃ©s para la aplicaciÃ³n. En general, la segmentaciÃ³n automÃ¡tica es una de las tareas mÃ¡s complicadas dentro del procesado de imagen. La segmentaciÃ³n va a dar lugar en Ãºltima instancia al Ã©xito o fallo del proceso de anÃ¡lisis. En la mayor parte de los casos, una buena segmentaciÃ³n darÃ¡ lugar a una soluciÃ³n correcta, por lo que, se debe poner todo el esfuerzo posible en la etapa de segmentaciÃ³n. Los algoritmos de segmentaciÃ³n de imagen generalmente se basan en dos propiedades bÃ¡sicas de los niveles de gris de la imagen: discontinuidad y similitud. Dentro de la primera categorÃ-a se intenta dividir la imagen basÃ¡ndonos en los cambios bruscos en el nivel de gris. Las Ã¡reas de interÃ©s en esta categorÃ-a son la detecciÃ³n de puntos, de lÃ-neas y de bordes en la imagen. Las Ã¡reas dentro de la segunda categorÃ-a estÃ¡n basadas en las tÃ©cnicas de umbrales, crecimiento de regiones, y tÃ©cnicas de divisiÃ³n y fusiÃ³n.

SegÃºn el grado de asociaciÃ³n entre las operaciones de segmentaciÃ³n y las de reconocimiento, se distinguen tres tipos principales de mÃ©todos de segmentaciÃ³n:

−Â Â Â Â Â Â Â Â Los MÃ©todos explÃ-citos oÂ segmentaciÃ³n en unidades fÃ-sicas: Estos mÃ©todos, intervienen avanzando el proceso de reconocimiento. Las partes segmentadas se dividen prÃ¡cticamente en letras, tanto que la segmentaciÃ³n se considera una parte del proceso de reconocimiento.

−Â Â Â Â Â Â Â Â Los MÃ©todos de segmentaciÃ³n implÃ-citos o segmentaciÃ³n en unidades lÃ³gicas: Los mÃ©todos implÃ-citos, consisten generalmente en una segmentaciÃ³n mÃ¡s fina y asÃ- conseguir los puntos de corte correctos. Las partes segmentadas son llamadas grafemas. Estos se usarÃ¡n mÃ¡s adelante, durante el proceso de reconocimiento/ clasificaciÃ³n. Los grafemas estarÃ¡n compuestos por fragmentos de caracteres, caracteres o grupos de caracteres.

−Â Â Â Â Â Â Â Â Los MÃ©todos de segmentaciÃ³n implÃ-citos y exhaustivos: En este caso, es el reconocimiento quien guÃ-a la segmentaciÃ³n, asÃ- que el sistema de evaluaciÃ³n que se aplica aquÃ- implica un reconocimiento por cÃ¡lculo de las posiciones sucesivas de la imagen y escoger las posiciones de

segmentaciÃ³n que se correspondan con las responsables de las partes mÃ¡s significativas.

4− Reconocimiento de caracteres: Esta es sin duda la etapa de mayor dedicaciÃ³n ya que el reconocimiento de caracteres es en resumen el proceso donde se aplica mÃ©todos de comparaciÃ³n de configuraciones a las formas de los caracteres leÃ-dos para determinar quÃ© caracteres alfanumÃ©ricos o signos de puntuaciÃ³n representan las formas. Debido a que la diversidad de tipos y formatos de letras (por ejemplo, negrita y cursiva) puede traducirse en grandes diferencias en la forma de los caracteres, el reconocimiento de caracteres puede dar errores.

5− Reconocimiento de fuentes.

6− VectorizaciÃ³n: este proceso transforma las caracterÃ-sticas de la imagen en una lÃ-nea poligonal o curvilÃ-nea, en resumen consiste en convertir imÃ¡genes que estÃ¡n formadas por pÃ-xeles en imÃ¡genes formadas por vectores. Esto se logra dibujando todos los contornos y rellenos de la imagen mediante curvas BÃ©zier, Las curvas BÃ©zier son ampliamente utilizadas en computaciÃ³n grÃ¡fica debido a que requieren poco espacio de almacenamiento y son independientes de la resoluciÃ³n de salida que se utilice. Su uso actual se extiende desde la representaciÃ³n de tipografÃ-as hasta el modelado de objetos tridimensionales. Los dibujos obtenidos mediante la vectorizaciÃ³n son imÃ¡genes de contornos perfectamente definidos, que pueden ampliarse o reducirse a cualquier tamaÃ±o sin que se modifique su alta calidad. (Ver Anexos, pÃ¡g. 15: Anexo No 1).

7− Reconocimiento de grÃ¡ficos (si es que los hay).

8− Reconocimiento estructural: determina la organizaciÃ³n lÃ³gica de las entidades elementales o compuestas, se asume una representaciÃ³n estructurada de los objetos, estando usualmente los subobjetos representados de manera similar a la utilizada en los mÃ©todos globales. Las formas se representan mediante una serie de "reglas de composiciÃ³n" que deben cumplir los subobjetos para pertenecer al conjunto, existiendo un amplio abanico de posibilidades (Ã¡rboles de decisiÃ³n, expresiones lÃ³gicas, redes, modelos de Markov, reglas, entre otros.

9− ClasificaciÃ³n de documentos: se distingue el tipo de documento reconocido.

En particular, en la etapa del reconocimiento de caracteres se divide en dos sub−etapas:

1− ExtracciÃ³n de caracterÃ-sticas:

− Permite conocer medidas (tamaÃ±o, perÃ-metro, centro de gravedad, momentos...).

− CaracterÃ-sticas topolÃ³gicas (orientaciÃ³n de segmentos, nÃºmero de agujeros, nÃºmero de extremidades, etc...).

2− Etapa de decisiÃ³n: Tres tÃ©cnicas destacan sobre las demÃ¡s:

− Redes neuronales (capacidad de aprendizaje).

− Cadenas Ocultas de Markov. Estudios y algoritmos probabilÃ-sticos.

− Voto mayoritario. CombinaciÃ³n de diferentes estrategias. Se escogerÃ¡ la clase con mayor nÃºmero de clasificaciones. Es la tÃ©cnica que mejores resultados ofrece.

En la mayorÃ-a de los tratamientos se requiere para su buen funcionamiento una contribuciÃ³n de informaciÃ³n del contexto, dependiente del tipo del documento analizado. Esta informaciÃ³n se proporciona

por los llamados modelos de documentos. Varias etapas del reconocimiento requieren esta clase de conocimientos: un reconocedor de caracteres utilizarÃ¡ por ejemplo una base de datos de caracteres de referencia o diccionarios lingÃ¼Ã-sticos; el reconocimiento de fuentes necesitarÃ¡ una base de conocimiento de las caracterÃ-sticas de las fuentes en cuestiÃ³n.

El reconocimiento de caracteres tiene como objeto el asociar a una imagen la identidad correspondiente de entre los sÃ-mbolos de un determinado alfabeto. Existen varias modalidades entre las cuales se pueden distinguir reconocimiento on line y reconocimiento off line, es decir teniendo o no en cuenta la informaciÃ³n temporal asociada a los trazos que componen el caracter. SÃ³lo se tratarÃ¡ aquÃ- la segunda opciÃ³n. Si la informaciÃ³n del caracter se obtiene a travÃ©s de medios Ã³pticos, suele hablarse de Reconocimiento Ã“ptico de Caracteres (OCR). En este caso, el problema del reconocimiento de los caracteres es sÃ³lo una parte de un problema mayor que se conoce como AnÃ¡lisis o ComprensiÃ³n de Documentos. Se trata, evidentemente, de obtener una representaciÃ³n simbÃ³lica lo mÃ¡s fiel y completa posible a partir de la imagen digitalizada de un documento escrito. Las etapas de que consta un sistema de anÃ¡lisis de documentos son: adquisiciÃ³n de la imagen y preproceso, segmentaciÃ³n de Ã©sta en bloques de grÃ¡ficos, bloques de texto, lÃ-neas de texto y caracteres, y ordenaciÃ³n de estos elementos, reconocimiento de los caracteres impresos independientemente del tipo y tamaÃ±o de letra y, finalmente, recuperaciÃ³n de errores en el texto por correcciÃ³n ortogrÃ¡fica y aplicaciÃ³n de un modelo de lenguaje.

La complejidad del sistema varÃ-a en funciÃ³n del tipo de caracteres que aparecen en el documento. En este sentido se distinguen las siguientes variedades de menor a mayor dificultad: reconocimiento de uno o varios tipos de letra impresa (fixed−font OCR y multifont OCR), de cualquier tipo de letra impresa (omnifont OCR), de caracteres aislados manuscritos (handwritten OCR) y, finalmente, de escritura manual cursiva (script recognition). En cuanto al mÃ©todo de reconocimiento propiamente dicho, dos aproximaciones distintas que coinciden con las que se distinguen en Reconocimiento de Formas en general son: el reconocimiento geomÃ©trico (estadÃ-stico o basado en la teorÃ-a de la decisiÃ³n) y el reconocimiento estructural. Cada uno de ellos tiene sus ventajas y sus inconvenientes. La informaciÃ³n estructural presente en los patrones a reconocer es difÃ-cilmente aprovechada, en general, por los mÃ©todos geomÃ©tricos. Esta informaciÃ³n, sin embargo, puede tener extraordinaria importancia en muchos casos. Por otro lado, los mÃ©todos de extracciÃ³n de primitivas (paso necesario previo a la aplicaciÃ³n de un mÃ©todo estructural y, en particular, sintÃ¡ctico) son generalmente mÃ¡s costosos y pueden causar problemas en presencia de fuerte ruido. Los propios mÃ©todos sintÃ¡cticos aÃ±aden a menudo mÃ¡s coste que los geomÃ©tricos y pueden ser tambiÃ©n mÃ¡s sensibles al ruido si no incorporan una formulaciÃ³n estocÃ¡stica. La mayor parte de los trabajos que abordan el reconocimiento de caracteres manuscritos hace uso de mÃ©todos hÃ-bridos sintÃ¡cticogeomÃ©tricos, en este trabajo se presenta un mÃ©todo de parametrizaciÃ³n, en vectores de talla fija, de imÃ¡genes digitalizadas de caracteres manuscritos. La representaciÃ³n de los objetos en forma de vectores es una caracterÃ-stica fundamental de un sistema de reconocimiento geomÃ©trico de formas. El espacio vectorial definido con esta representaciÃ³n es el marco de aplicaciÃ³n de gran cantidad de tÃ©cnicas tanto paramÃ©tricas como no paramÃ©tricas. Entre las primeras pueden citarse las basadas en la estimaciÃ³n de probabilidad asumiendo distribuciones normales, mezclas de gaussianas, etc. Entre las segundas, las funciones discriminantes, la bÃºsqueda de vecinos y las ventanas de Parzen asumiendo una mÃ©trica, etc. Tanto unas como otras se apoyan en bases teÃ³ricas procedentes de la estadÃ-stica y la teorÃ-a de la decisiÃ³n.

A continuaciÃ³n se presenta de forma resumida el mÃ©todo de ParametrizaciÃ³n:

ParametrizaciÃ³n: El objetivo es encontrar un conjunto de parÃ¡metros que definan la forma del caracter bajo anÃ¡lisis de forma precisa y Ãºnica. AdemÃ¡s, la continuidad de la representaciÃ³n es imprescindible para asegurar el mayor grado de inmunidad al ruido y la mayor capacidad de generalizaciÃ³n posibles. Esto significa que objetos similares deben dar lugar a representaciones similares. En conclusiÃ³n, tres caracterÃ-sticas que definen un mÃ©todo de parametrizaciÃ³n adecuado son: precisiÃ³n, unicidad y continuidad. Por otro lado la concisiÃ³n del mÃ©todo de parametrizaciÃ³n, dependiente del nÃºmero de

parÃ¡metros que se generan (dimensionalidad de la salida) y en menor medida del rango y la cuantizaciÃ³n de Ã©stos, es tambiÃ©n un factor clave por varios motivos. Uno de ellos es la complejidad temporal y espacial de muchos algoritmos de clasificaciÃ³n, que crece muy deprisa cuando aumenta la dimensionalidad de los vectores a clasificar. En algunos casos este crecimiento puede llegar a ser exponencial. Otro argumento en contra de una alta dimensionalidad de la representaciÃ³n es el incremento exponencial de la dispersiÃ³n de un conjunto de puntos en espacios de dimensiÃ³n creciente. Esto significa que la obtenciÃ³n de estimaciones estadÃ-sticamente fiables en estos espacios requiere un nÃºmero de muestras que crece exponencialmente con la dimensiÃ³n, en realidad es el aumento de la dimensionalidad intrÃ-nseca o topolÃ³gica la que determina este hecho, pero Ã©sta crece generalmente con la dimensionalidad algebraica. Otros aspectos tambiÃ©n relevantes a la hora de diseÃ±ar o escoger un mÃ©todo de parametrizaciÃ³n dependen directamente de las particularidades del problema. El factor quizÃ¡ mÃ¡s importante a este respecto es la invarianza de la representaciÃ³n. La parametrizaciÃ³n debe ser invariante a distorsiones o alteraciones que no varÃ-en la naturaleza del objeto representado, esto es, que no hagan variar la clase en la que Ã©ste se va a clasificar finalmente. Desgraciadamente, esta importante propiedad es difÃ-cil de obtener a travÃ©s de mÃ©todos analÃ-ticos o teÃ³ricos y se aborda generalmente mediante heurÃ-sticos y procedimientos hasta cierto punto intuitivos.

ClasificaciÃ³n geomÃ©trica de patrones: Se han empleado diversos mÃ©todos de clasificaciÃ³n para evaluar las prestaciones del mÃ©todo de representaciÃ³n LLF (Local Line Fitting). Una red neuronal (PerceptrÃ³n Multicapa), una variante del LVQ (Learning Vector Quantization) llamada DSM (Decision Surface Mapping), la regla de clasificaciÃ³n del vecino mÃ¡s prÃ³ximo con la (distancia euclÃ-dea)* como medida de distancia y esta misma regla con un conjunto de prototipos editado y condensado han sido los paradigmas utilizados. Las redes neuronales son estructuras formadas por elementos de proceso muy simples. La topologÃ-a de un perceptrÃ³n multicapa consiste en cierto nÃºmero de capas de unidades conectadas en cascada. A la primera capa (capa de entrada) se le presentan las muestras a clasificar (vectores de dimensiÃ³n fija). Las capas posteriores (capas ocultas) alimentan la capa de salida para que Ã©sta proporcione una salida (otro vector de dimensiÃ³n fija) en funciÃ³n de los pesos de los enlaces entre unidades. Son precisamente estos pesos los que se van adaptando en la fase de aprendizaje por medio de algoritmos que comparan el resultado proporcionado por la red con el deseado y modifican el valor de los pesos de modo que la discrepancia sea mÃ-nima. En este caso se utiliza el algoritmo Backward error propagation sobre un perceptrÃ³n con tres capas de pesos.

La regla del vecino mÃ¡s prÃ³ximo es una de las tÃ©cnicas de clasificaciÃ³n mÃ¡s conocidas. Sea *{x1, x2, ..., xn}* un conjunto de *n* muestras etiquetadas. Sea *xi*. *X* la muestra mÃ¡s cercana a *x* segÃºn una cierta medida de distancia. La aplicaciÃ³n de la regla del vecino mÃ¡s prÃ³ximo supone asignar a *x* la clase a que pertenece *xi.* El anÃ¡lisis para *n*Ã† demuestra que el error cometido nunca es mayor que el doble del error de Bayes para un conjunto infinito de muestras. La regla de los *k*−vecinos mÃ¡s prÃ³ximos es una extensiÃ³n de la anterior en la que *x* se clasifica en la clase que mÃ¡s veces aparezca en el conjunto de los prototipos *xi1,xi2...xiK* mÃ¡s cercanos a *x.* El nÃºmero Ã³ptimo de vecinos depende fuertemente (en relaciÃ³n inversa) de *n* y de la dimensiÃ³n del espacio en que se definen los patrones. Aunque la dimensiÃ³n aplicable es la topolÃ³gica, una estimaciÃ³n probablemente optimista (asumiendo incluso una dimensiÃ³n topolÃ³gica un orden de magnitud menor que la algebraica) para nuestro problema revela una *k* Ã³ptima menor que *1.* Este hecho se ha comprobado experimentalmente en el caso de los dÃ-gitos con LLF donde los mejores resultados se obtienen con *k=1.* Las tÃ©cnicas de ediciÃ³n y condensado son ya clÃ¡sicas en el Ã¡rea de reconocimiento de formas. El objetivo es obtener un subconjunto del conjunto de muestras original que preserve en la medida de lo posible las fronteras de decisiÃ³n de Ã©ste. Un menor nÃºmero de prototipos implica un reconocimiento mÃ¡s eficiente y, en algunos casos, mejores prestaciones debido a la presencia masiva de outliers en el conjunto de entrenamiento. Otro tipo de tÃ©cnicas son las de modificaciÃ³n adaptativa de los prototipos. Las mÃ¡s conocidas son, posiblemente, las tÃ©cnicas LVQ. En este caso, se da libertad a los prototipos para moverse libremente en el espacio de caracterÃ-sticas. Es evidente que esta libertad no puede existir si el espacio en el que se definen las muestras no es vectorial (bien sea mÃ©trico, pseudomÃ©trico, etc.). La aproximaciÃ³n adoptada ha sido, concretamente, llamada allÃ- DSM (Decision Surface Mapping). La caracterÃ-stica principal de esta tÃ©cnica es la modificaciÃ³n del prototipo mÃ¡s

cercano de la misma clase que la muestra acercÃ¡ndolo a Ã©sta y la modificaciÃ³n del mÃ¡s cercano de otra clase alejÃ¡ndolo de ella. Las muestras sÃ³lo provocan la alteraciÃ³n de prototipos si son clasificadas incorrectamente con el conjunto de prototipos actual. Aunque las superficies de decisiÃ³n obtenidas no respetan (en la misma medida que las tÃ©cnicas LVQ) las fronteras de decisiÃ³n de Bayes, los resultados experimentales parecen superar ampliamente los obtenidos con LVQ. Se utiliza una variante original de la aproximaciÃ³n DSM en la que se fija el error mÃ¡ximo de resubstituciÃ³n y no el nÃºmero de prototipos. De esta forma, la especificaciÃ³n del modelo es mucho mÃ¡s natural. El coste temporal es mayor, sin embargo, ya que el mÃ©todo empieza por asignar un prototipo a cada clase y va aÃ±adiendo nuevos prototipos a aquellas clases que no superen el error de resubstituciÃ³n prefijado.

CONCLUSIONES

Debido al problema que ocurre en la actualidad en cuanto a la conservaciÃ³n de los documentos en soporte papel ante todo cuando tiene hasta un siglo de antigÃ¼edad, al igual que la informaciÃ³n en las bibliotecas, que se esta trabajando para que sea digital debido a que ya existen muchos de sus libros deteriorados, ademÃ¡s de que por una causa lÃ³gica debemos ir a favor del desarrollo y por supuesto es una forma mas cÃ³moda, menos costosa, mas fÃ¡cil y mas eficiente para el trabajo, la ciencia se ha esmerado por trabajar dedicadamente en diseÃ±ar mÃ©todos para la digitalizaciÃ³n de los mismos. Todos conocemos que por medio de un escÃ¡ner se puede digitalizar libros, revistas, documentos, cartas, telegramas, entre otros; como tambiÃ©n ocurre el proceso de digitalizaciÃ³n a travÃ©s de las cÃ¡maras digitales, las minicomputadoras portÃ¡tiles, la nueva telefonÃ-a movil, la tecnologÃ-a de Ipod reciente, pero falta un paso muy importante que consiste en interpretar el contenido e informaciÃ³n de estas imÃ¡genes digitales. En la actualidad existe gran variedad de software para el reconocimiento de caracteres pero sin embargo poseen una margen de error alto cuando la entrada es un documento de caracteres manuscritos. Debido a esto se realizÃ³ este trabajo en aras de conocer todo el procedimiento principal que utiliza el Reconocimiento de Caracteres Manuscritos mientras digitaliza la imagen, ademÃ¡s de que ya existen diversos mÃ©todos y sistemas para este proceso aunque aun no se ha logrado demasiada perfecciÃ³n en la practica de los mismos. TambiÃ©n se desarrolla uno de estos mÃ©todos que fue en este caso el que se escogiÃ³ para explicar, pues fue el que despertÃ³ mayor interÃ©s y se conoce como El Sistema Ã“ptico de Caracteres Manuscritos que estÃ¡ basado en el uso de un dispositivo de exploraciÃ³n Ã³ptica que puede reconocer la letra impresa.

La tÃ©cnica del Reconocimiento de Caracteres Manuscritos comenzÃ³ aplicÃ¡ndose en documentos para los cuales ninguna forma electrÃ³nica estaba disponible y a medida que la tecnologÃ-a ha avanzado, sus aplicaciones han evolucionado. El reconocimiento de caracteres sigue siendo un problema complejo que tropieza con dificultades aÃºn no resueltas y que son actualmente un objeto de numerosas investigaciones.

RECOMENDACIONES

DespuÃ©s de haberse investigado e indagado en el tema del Reconocimiento de Caracteres Manuscritos se puede decir que:

- SerÃ-a bueno que la ciencia trabajara mÃ¡s profundo aÃºn, en el tema, debido a que las tÃ©cnicas del Reconocimiento de Caracteres Manuscritos son por lo general muy costosas y no estarÃ-a de mÃ¡s proponerse un menor gasto de recursos.
- Seguir trabajando en el perfeccionamiento de las tÃ©cnicas y mÃ©todos utilizados en el Reconocimiento de Caracteres Manuscrito ya que distan mucho de ser perfectos y aÃºn en la actualidad poseen un margen de error muy grande.
- Que cada dÃ-a la ciencia innove mÃ¡s dispositivos en los cuales se pone de manifiesto el Reconocimiento de Caracteres Manuscritos como son la nueva telefonÃ-a mÃ³vil, las computadoras de bolsillo que son con pantalla tÃ¡ctil, la Ãºltima tecnologÃ-a de Ipod, las cÃ¡maras digitales, escaners, fotocopiadoras es decir todo aquello que cada dÃ-a nos facilita mÃ¡s el trabajo, la comunicaciÃ³n, como tambiÃ©n el despeje.

BIBLIOGRAFIA

Blasco LÃ³pez, Antonio y FÃ©lez Esteban, Francisco: Reconocimiento de Caracteres Manuscritos. sl, se, sf.

PÃ©rez, Juan Carlos y Vidal, Enrique y SÃ¡nchez, Lourdes: Sistema GeomÃ©trico de Reconocimiento Ã“ptico de Caracteres. Valencia, se, sf.

BÃ¡ez Rojas, J.J y Guerrero, M.L y Conde Acevedo, J y Padilla Vivanco, A y Urcid Serrano, G: SegmentaciÃ³n de ImÃ¡genes. MÃ©xico, se, 2004.

Martin, Marcos: TÃ©cnicas ClÃ¡sicas de SegmentaciÃ³n de ImÃ¡genes. sl, se, 2002.

Rojas, Javier: SegmentaciÃ³n y Reconocimiento de Patrones. sl, se, 2005.

Toscano Medina, Karina y Nakano Miyatake, Mariko y SÃ¡nchez PÃ©rez, Gabriel y PÃ©rez Meana, HÃ©ctor M y Yasuhara, Makoto: Reconocimiento de Caracteres Manuscritos, en Instituto PolitÃ©cnico Nacional, CientÃ-fica, MÃ©xico, 9(003), 2005, p. 143−154.

ANEXOS

Anexo No 1: [1]

AquÃ- van algunos ejemplos de las curvas BÃ©zier:

Curvas lineales de BÃ©zier

Dado los puntos P0 y P1, una curva lineal de BÃ©zier es una lÃ-nea recta entre los dos puntos. La curva viene dada por la expresiÃ³n:

B(t) = P0 + (P1 − P0)t = (1 − t)P0 + tP1, t Ð„ [0,1].

Curvas CuadrÃ¡ticas de Bezier

Una curva cuadrÃ¡tica de BÃ©zier es el camino trazado por la funciÃ³n B(t), dados los puntos: P0, P1 y P2,

B(t) = (1 − t)Â² P0 + 2t(1 − t)P1 + tÂ²P2, t Ð„ [0,1].

Curvas CÃºbicas de Bezier

Cuatro puntos del plano o del espacio tridimensional, P0, P1, P2 y P3 definen una curva cubica de BÃ©zier. La curva comienza en el punto P0 y se dirige hacia P1 y llega a P3 viniendo de la direcciÃ³n del punto P2. Usualmente no pasara por P1 ni por P2. Estos puntos solo estÃ¡n ahÃ- para proporcionar informaciÃ³n direccional. La distancia entre P0 y P1 determina que longitud tiene la curva cuando se mueve hacia la direcciÃ³n de P2 antes de dirigirse hacia P3.

La forma parametrica de la curva es:

B(t) = P0(1 − t)Â³ + 3P1t(1 − t)Â² + 3P2tÂ²(1 − t) + P3tÂ³, t Ð„ [0,1].

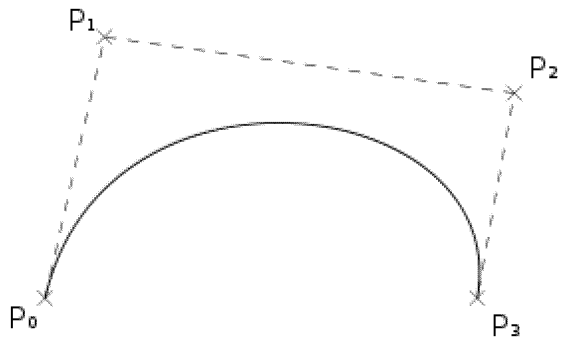

Fig. No 1: Curva CÃºbica de BÃ©zier donde se aprecian los puntos o nodos* de anclaje P1 y P2.

[1] Paul Bourke: BÃ©zier curves, http://local.wasp.uwa.edu.au/~pbourke/

GLOSARIO

En esta secciÃ³n va la definiciÃ³n de toda aquella palabra que fue marcada con * porque se entendiÃ³ que podÃ-an causar dificultades al conocer su significado.

- EntropÃ-a: Magnitud termodinÃ¡mica que indica el grado de desorden molecular de un sistema.
- Distancia EuclÃ-dea: es la distancia normal que todos conocemos.
- Nodos: es un punto de interseccciÃ³n o uniÃ³n de varios elementos que confluyen en el mismo lugar

13