1.1 Introducción.

El problema de la mayoría de las personas e instituciones, es la conservación de sus documentos escritos, especialmente cuando éstos tienen un siglo de antigüedad. Es por esto que se ha intentado conservar tanto el documento, como la información que éste contiene.

Un avance en la tecnología a finales del siglo XX, ha dado origen al desarrollo de las bibliotecas digitales, las cuales ponen a disposición de un mayor número de personas la información. Por medio de un escáner se puede digitalizar libros, revistas, documentos (cartas, telegramas, memorandas,...); pero falta un paso muy importante, interpretar el contenido e información de estas imágenes digitales.

Existen en el mercado una variedad de software de reconocimiento de caracteres (OmniPage [OmniPage 03], Corel [Corel 03], Proyecto Clara OCR [CLARA 03], …), sin embargo todos poseen un margen de error muy alto cuando la entrada es un documento con caracteres manuscritos. Es por ello que se han propuesto proyectos de investigación acerca del reconocimiento de caracteres manuscritos. La Universidad de las Américas - Puebla (UDLAP), tiene actualmente un Proyecto de Investigación dedicado al reconocimiento de caracteres manuscritos, aplicado a los telegramas escritos por el *Gral. Porfirio Díaz*, dado que es una contribución importante a la historia del País [Gómez, Linarez, Spínola y Cortés 01]. Se pretende eventualmente digitalizar los telegramas del *Gral. Porfirio Díaz*, que la Biblioteca de la Universidad de las Américas – Puebla actualmente maneja, estos se encuentran en microfilm, transcribiendo el texto que contiene. La figura 1.1 muestra la descripción general del sistema. El proyecto descrito aquí continua las investigaciones realizadas por Sergio Linares y Carlos Alberto Spínola [Linares & Spínola 00], plasmadas en la tesis "*Reconocimiento de Letra Manuscrita de Porfirio Díaz, Utilizando un Shell*

Neuronal ANNSYD" y los estudios de Jorge Navarrete reportados en su tesis "*Mejora del Algoritmo de segmentación para el reconocimiento de caracteres de telegramas escritos por el Gral. Porfirio Díaz*" [Navarrete 02]. El trabajo de [Linares & Spínola 00] encontró una solución básica a este problema, pero no obtuvo resultados suficientemente buenas para su implementación practica. Actualmente se ha divido al proyecto en dos partes: mejoras a la segmentación de letras en una palabra y mejoras al reconocimiento de caracteres. Se aplicarán técnicas diferentes que mejoren los resultados en cada parte. A la fecha en el caso de la segmentación se logró un 83.86% de éxito [Navarrete 02], coeficiente que puede ser mejorado con un archivo de entrenamiento mayor, ya que sólo se usaron 80 palabras.

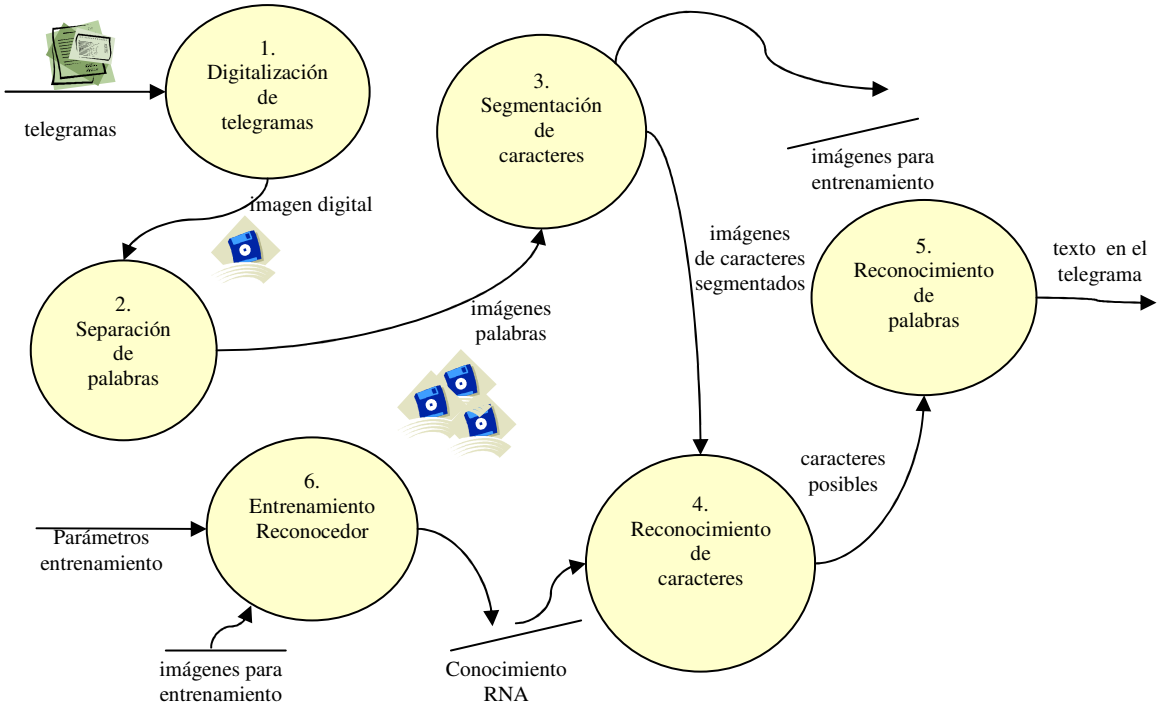

Figura 1.1 Descripción General del Proyecto [Gómez 03].

Una buena segmentación 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 aún para el ser humano. En el caso de los telegramas del Gral. Porfirio Díaz cabe añadir, que dado su antigüedad (alrededor de un siglo) la escritura y el vocabulario utilizado han sufrido modificaciones a lo largo del tiempo al compararse con la escritura de la actualidad (ver figura 1.2).

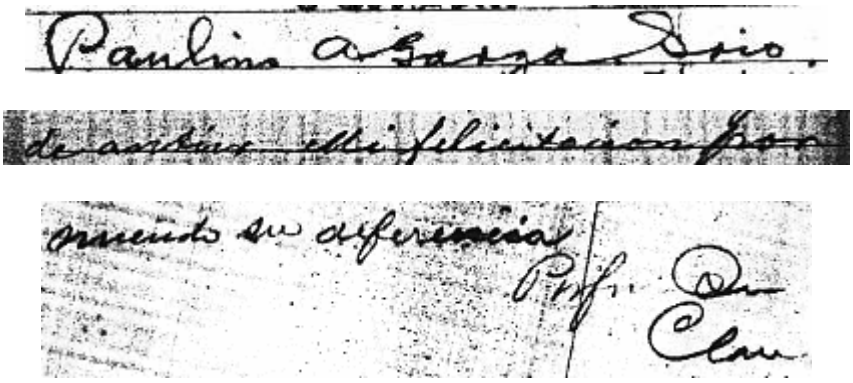

Figura 1.2. Ejemplo de palabras tomadas de los telegramas del Gral. Porfirio Díaz.

En esta investigación se continua con el siguiente eslabón (Burbuja 4 figura 1.1), que es el de reconocer cada caracter de la palabra segmentada. Para esto se realizan pruebas con dos algoritmos de clasificación: Vecino Más Cercano usando la distancia euclidiana y una Red Neuronal de *Kohonen*. Cabe mencionar que esta investigación esta en el campo de reconocimiento de patrones.

1.2 Reconocimiento de Caracteres Ópticos.

El Reconocimiento de Caracteres Óptico (OCR, Optical Character Recognition), comenzó desde 1951 con al invención de M. Sheppard's, de un robot lector – escritor. En 1954 J. Rainbow desarrolló un prototipo de una máquina que podía leer a la fantástica velocidad de un carácter en mayúscula por minuto. Varias compañías, incluyendo IBM, Recognition

Equipment, Inc., Farrington, Control Data, y Optical Scanning Corporation, comercializaron Software de OCR en los años 1967 [Srihari & Lam 92].

El proceso de reconocimiento empieza desde la digitalización del documento (ver figura 1.3), esto se realiza a través de un escáner, a continuación se describen las fases del reconocimiento:

- Análisis del documento. Esto consiste en la extracción de los caracteres a reconocer. Una buena segmentación del caracter y reconocimiento del mismo dependen de la calidad de la imagen digitalizada.

  Antes de hacer el reconocimiento del caracter es necesario aislar los caracteres individuales. En reconocimiento de texto escrito a mano esto puede ser muy difícil, por que los caracteres pueden estar sobrepuestos, y el tamaño de un mismo caracter puede variar, así como la inclinación y el grosor de éste.

- Reconocimiento del caracter. Se realiza por medio de algoritmos de reconocimiento de patrones. Aquí se extraen las características principales del caracter los cuales son usados para su clasificación.

- Procesamiento Contextual. Consiste en utilizar la información que nos proporciona el reconocedor de caracter, ya que esto es se puede enlazar al dato obtenido para localizar más información relacionada con ésta. Veamos un ejemplo: en los Estados Unidos de América, el Servicio Postal implementó un sistema de reconocimiento del código postal, el sistema localiza el bloque de la dirección del destino, de la cual

extrae el código postal, el sistema genera un código de 9 dígitos, que es utilizado para ordenarlos [Srihari & Lam 92].

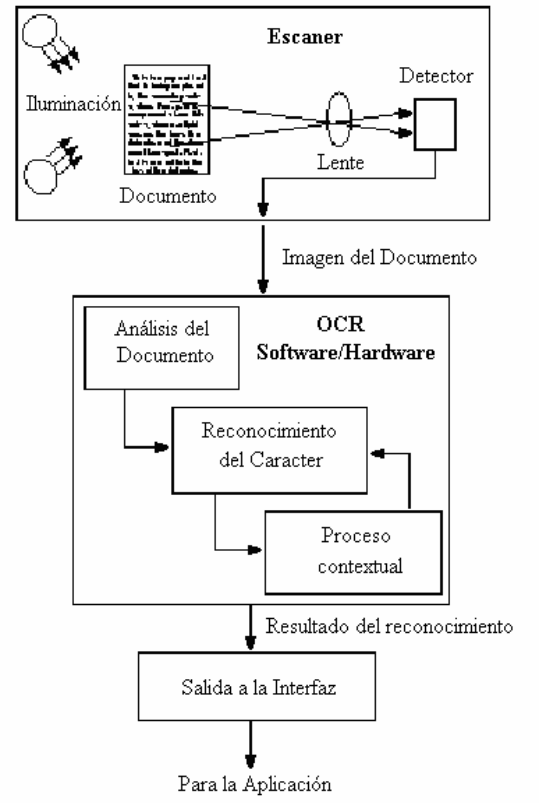

Figura 1.3. Diagrama de cómo se realiza la digitalización de un documento [Srihari & Lam 92].

1.3 Reconocimiento de Patrones.

Este trabajo esta dentro del reconocimiento de patrones, por lo cual se definen ciertos conceptos.

- Clase ó Patrón: Conjunto de entidades que comparten alguna característica que las diferencia de otras.
- Extractor de características: subsistema que extrae información relevante para la clasificación a partir de las entidades cuantificables

- Clasificador: subsistema que utiliza un vector de características de la entidad cuantificable y lo asigna a una de M clases [Tou & Gonzalez 81] (ver figura 1.4).

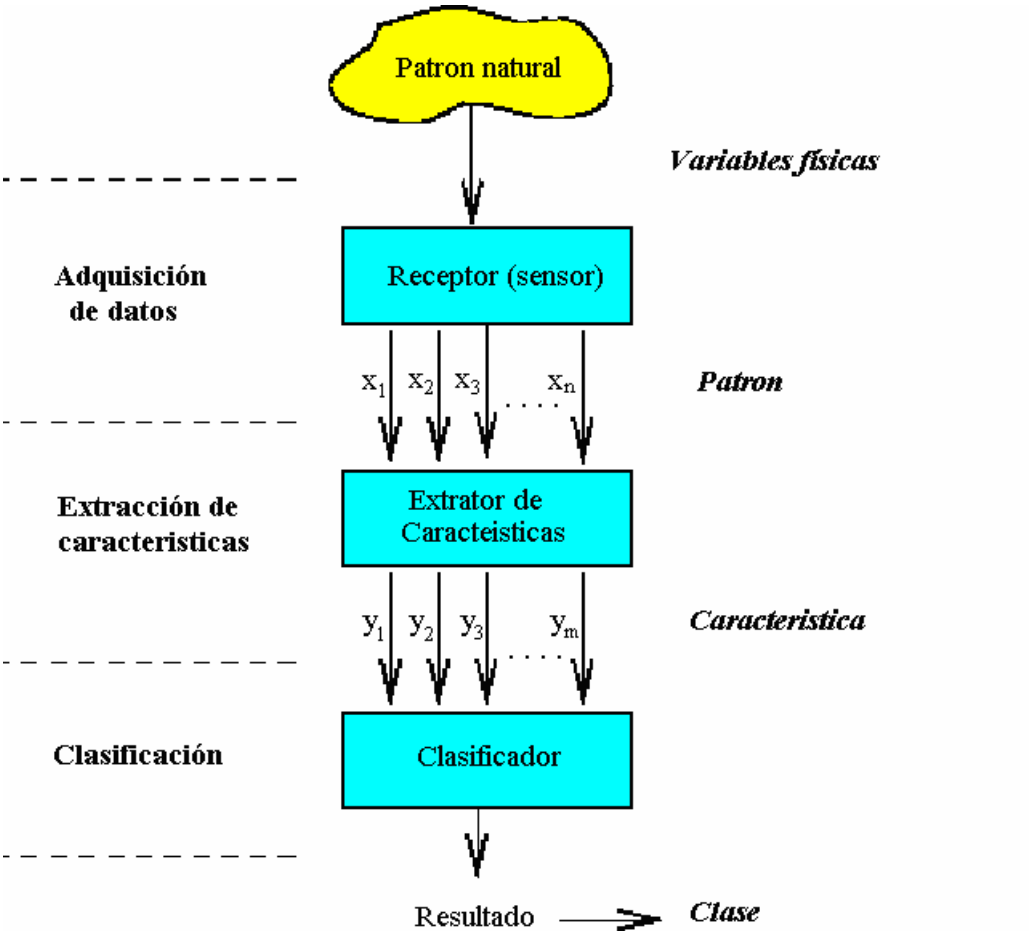

Figura 1.4. Etapas de un sistema de reconocimiento de patrones [Cortijo 01].

En [Gómez & Oldham 93] se 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.

1.4 Trabajos Relacionados.

En seguida se comentan algunos trabajos relacionados al reconocimiento de caracteres manuscritos. Cabe mencionar que existe una amplia bibliografía relacionada con el reconocimiento de caracteres, reconocimiento de caracteres en línea (para las computadoras Pocket PC), y reconocimiento de caracteres manuscritos.

Reconocimiento Automático de Matrículas.

En el laboratorio de la Facultad de Informática de la Universidad Politécnica de Madrid, España, se desarrolló para una empresa privada entre los años 1990 y 1991, un sistema automático para el reconocimiento de matrículas de vehículos industriales. Este proyecto dío como resultado una alta efectividad en el reconocimiento, obteniendo alrededor de 90% de reconocimiento. Mencionan que se puede llegar a un 100% de reconocimiento si las matrículas se encuentran en buen estado. Cabe mencionar que como entrada para el reconocedor se da una matriz de puntos de 32 x 16. [Maravall 93].

Modelos Neuronales para Letras Cursivas Escritas a Mano.

En esta investigación el autor hace una comparación de la una red de tres capas de perceptrones y una red de *Kohonen*, no muestra porcentaje de reconocimiento, solo menciona que la red de *Kohonen* proporciona un mejor método para la codificación eficiente y mas compacto. La red generó un mapa grafotópico como el que se ve en la

figura 1.5 correspondiente a un entrenamiento con 15 x15 = 225 neuronas de salida y 10 de entrada [Morasso 88].

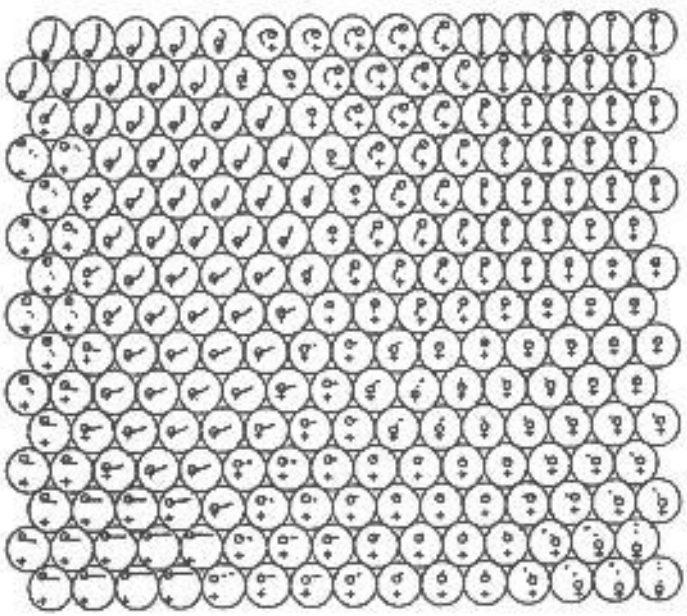

Figura 1.5 Visualización del mapa grafotópico [Morasso 88].

Reconocimiento del Texto Manuscrito

[Pittman 91] implementa una red de retropropagación con 2 capas escondidas, cada capa usa una interconexión global con la capa anterior. El autor entrenó la red con 50,000 letras cursivas. Tuvo un reconocimiento de 73% en el entrenamiento y 63% durante pruebas generales.

Reconocimiento de Caracteres Manuscritos con Redes Recurrentes Neuronales.

Esta es una tesis de la Universidad Politécnica de Valencia, en donde el objetivo del proyecto fue la construcción de un sistema de reconocimiento de caracteres manuscritos con redes recurrentes. En particular verificaron la potencia de las redes recurrentes para el

reconocimiento de caracteres manuscritos aislados, en la cual tuvieron un porcentaje de reconocimiento de un 70 a 80%. [Faraón 01].

Reconocimiento de Firmas usando un Arreglo de Perceptrones Multicapas.

El sistema propuesto por [Toscazo, Sánchez, Nakano y Pérez 00] consiste en una etapa de extracción de envolventes y otra de la extracción de parámetros estadísticos. Seguido del entrenamiento en paralelo de cinco redes neuronales tipo retropropagación, cada una de las cuales se entrenarán con los valores obtenidos de las envolventes superior, inferior, derecha e izquierda de la firma y la quinta red se entrenará con los valores obtenidos de los parámetros estadísticos. Para la evaluación del sistema, generaron una base de datos la cual consistió en 130 firmas (15 firmas de 9 personas), 90 firmas (10 firmas de 9 personas) que fueron usadas para entrenar cinco redes separadamente, 45 firmas (5 firmas de 9 personas) fueron usadas para la evaluación de funcionamiento del sistema. Los resultados de la simulación computacional fueron realizados en una Workstation Kayak XA y mostraron un reconocimiento aproximado del 100 por ciento de las firmas.

Segmentación y Reconocimiento de Fechas Manuscritas.

En este articulo presentan un sistema híbrido de los Modelos Ocultos de Markov y una Red Neuronal Multi-Capa, para reconocer las fechas escritas en los cheques del Banco de Brasilia. La red fue usada para el reconocimiento de los dígitos y los Modelos de Markov para el reconocimiento y validación del mes. El sistema dío un 95.5% de reconocimiento de la fecha, con un conjunto de prueba compuesto por 400 imágenes [Morita, Sabourin, Bortolozzi, y Suen 02].

1.5 Objetivo General.

Este proyecto consiste en analizar dos algoritmos para el reconocimiento de caracteres manuscritos antiguos.

Para esto, se realizará un software de reconocimiento de caracteres, en donde se implementan dos algoritmos: una red neuronal basada en la arquitectura de Kohonen y el algoritmo de "el Vecino Más Cercano"

1.6 Objetivos Específicos.

- Implementar una Red Neuronal de *Kohonen*. Se prueban diferente arquitecturas para encontrar el mejor resultado de reconocimiento.
- Implementar un Algoritmo de tipo *Nearest Neighbors* (Vecino más cercano). Se analizarán los diferentes algoritmos, y se implementará el que proporcione un patrón de clasificación óptimo a nuestro problema.
- Probar nuevamente, los casos evaluados anteriormente por Sergio Linares y Carlos Alberto Spínola [Linares & Spínola 00], con los algoritmos de *Kohonen* y *Nearest Neighbors*.
- Comparar el índice de reconocimiento obtenido en las investigaciones de [Linares & Spínola 00], contra el generado por la red neuronal de *Kohonen* y el algoritmo de *Nearest Neighbors*.

1.7 Hardware y Software Utilizado.

Se cuenta con el siguiente equipo de Cómputo:

- Desktop Ensamblada, Procesador Pentium III a 1 GHz., Memoria 1Gb en RAM, Sistema Operativo Windows XP Home Edition.

Lenguaje de desarrollo C++Builder.

1.8 Descripción del Documento.

El capítulo dos contiene una breve introducción al tema de redes neuronales, así como la descripción de la Red de *Kohonen*, su algoritmo de entrenamiento, y algunas aplicaciones relacionadas con el área de reconocimiento de patrones. También se describe la técnica de agrupamiento *K-Means*, así como que es el reconocimiento basado en el Vecino Más Cercano por medio de la distancia Euclidiana.

En el capitulo tres se describen los programas que sirvieron para entender la clasificación respecto a la distancia Euclidiana, *K-Means*, y la red de *Kohonen*. Se describe el diseño de la aplicación de Reconocimiento de Letras Manuscritas.

El capitulo cuarto se dan las resultados de las pruebas realizadas para el reconocimiento de letras manuscritas, tanto con la Red de *Kohonen* como con Vecino Más Cercano.