Universitat Politècnica de València

Departament d'Informàtica de Sistemes i Computadors

Proyecto Final de Carrera

17 de Junio de 2011

Un sistema automático de identificación de tipos de documentos

Autor: Vicente Castelló Fos
Director: Joaquim Francesc Arlandis Navarro
Titulación: Ingeniería en informática

Índice general

Capítulo 1

Introducción

1.1. Tarea a resolver

En la sociedad actual una empresa puede recibir o generar diariamente una cantidad de documentos en papel que deben ser organizados y archivados. Con un escáner de alta velocidad, consumiendo poco tiempo y esfuerzo, esta empresa puede convertir los documentos recibidos a un formato digital, ante esto surge la necesidad de organizarlos y agruparlos en tipos similares para facilitar posteriores tratamientos.

Una tarea habitual en la que se da un problema específico de clasificación de documentos es la organización de recibos, formularios, facturas, documentos legales, informes médicos o documentos administrativos para su posterior proceso (p.ej. OCR) y/o almacenamiento. Siempre que estos documentos puedan ser divididos en categorías tales como "Factura del proveedor X", "Formulario de encuesta tipo Y", etc. se podría aplicar algún método típico de clasificación. Pero en este caso, solo parte de los documentos de cada clase permanecen invariables, mientras que el resto es distinto en cada instancia. La parte común puede ser significativamente menor que la parte variable, que puede estar compuesta por gran cantidad de texto manuscrito o impreso.

Así, un documento perteneciente al ciclo de trabajo descrito puede ser visto como una imagen con contenidos estáticos (fijos o preimpresos) y variables (impresos a maquina, manuscritos, pegados, estampados, etc.). Bajo esta definición, una categoria, tipo o clase de documentos se define como un conjunto de imágenes con la misma información estática, y diferente de la del resto de las clases. La información variable puede ser distinta dentro de los documentos de una misma clase, tanto en tamaño como en contenido, como se ha apuntado anteriormente. En la figura 1.1, se muestran ejemplos de algunos tipos de documentos con información estática y variable.

Este trabajo se enfrenta a la tarea de identificación de imágenes de documentos que provienen de múltiples dominios de aplicación, sin tener en cuenta su diseño, estructura, contenidos textuales y no textuales, así como diferentes cantidades de contenido cumplimentado [1].

[1] En este documento se utilizarán las expresiones "información cumplimentada" o "contenido rellenado" para referirse al término inglés *"filled-in content"*. Con ellas se hará referencia a la parte de un documento que no es constante en todas las muestras del mismo tipo. P. ej.: las respuestas del usuario en un formulario o encuesta, o el detalle de una factura.

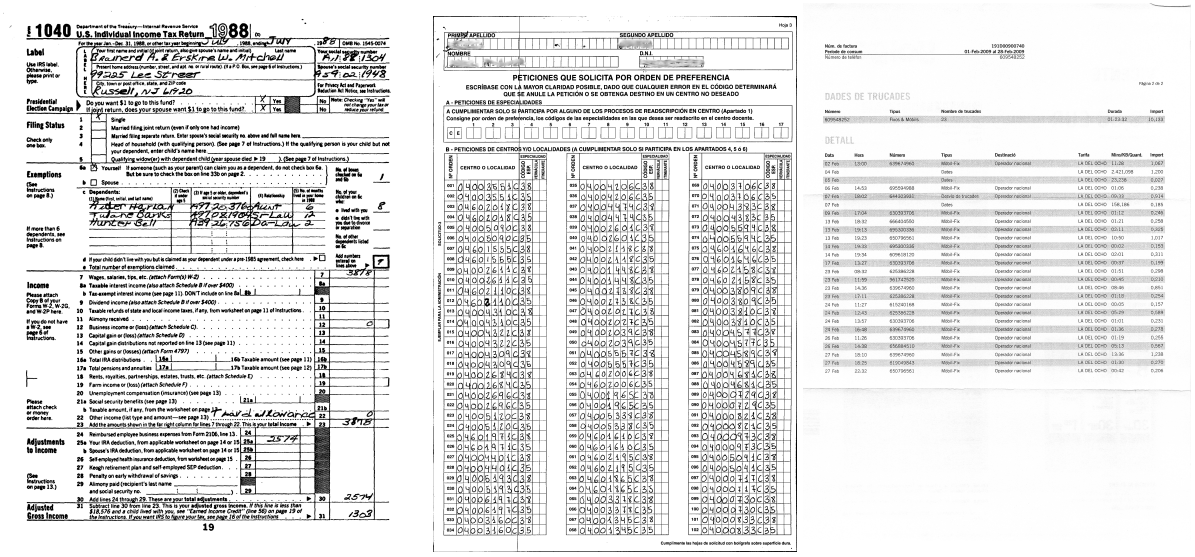


Figura 1.1: Ejemplos de documentos. El primero es un formulario en el que la información estática abarca la mayor parte del documento. El segundo es una página de formulario con gran cantidad de celdas que pueden estar cumplimentadas o no. El tercero es un documento con pocos elementos estructurales y contenidos estáticos, pero con mucha información variable.

1.2. Aproximación científica empleada

El presente trabajo está enmarcado en el campo del reconocimiento de formas y análisis de imágenes, y se centra en el estudio y experimentación de una metodología de clasificación para la identificación automática supervisada de imágenes de documentos aplicando una técnica basada en la extracción de características locales mediante un clasificador de los k-vecinos más próximos. El trabajo está acompañado de una experimentación exhaustiva para la identificación y optimización de los parámetros más relevantes y la evaluación de las prestaciones del sistema.

El problema práctico a resolver es la identificación de cualquier tipo de documento a partir de su imagen digital. Dada una serie de tipos de documentos conocidos (clases) representados por una o más imágenes de referencia, el sistema de identificación automática debe identificar una nueva imagen de test asignándola a una de las clases conocidas, o rechazándola si la imagen no pertenece a ninguna de ellas. Se pretende que el sistema sea capaz de identificar una imagen entre decenas o centenares de clases de documentos.

Un objetivo específico es conseguir que el sistema presente índices de identificación satisfactorios frente a imágenes de test que contengan información añadida respecto de la imagen de referencia de su clase. Este es el caso de los documentos cumplimentados con información manuscrita, por ejemplo formularios, o también rellenados con texto impreso, como ocurre, por ejemplo, con las facturas.

Para la resolución del problema planteado se seguirá una aproximación inductiva. Esta se aplica a problemas de reconocimiento de formas para los que no se encuentra una explicación satisfactoria o tangible sobre cuáles son los pasos o mecanismos a seguir para alcanzar una solución. Se realizará un aprendizaje supervisado, partiendo de un conjunto de entrenamiento etiquetado con c clases predeterminadas con el que se entrenará un clasificador para que frente a cualquier observación de test (no vista anteriormente) se asigne la clase más probable. Para la clasificación se empleará una metodología estadística (o geométrica), basada en la teoría de la decisión, donde un objeto está representado por d

características y es tratado como un punto de un espacio vectorial *d*-dimensional. La clasificación de un objeto en una determinada clase se decide en función de su posición en este espacio y de las distancias que lo separan de los otros puntos.

Hay muchas reglas de clasificación. En este caso, el clasificador elegido es una combinación del de los *k*-vecinos más próximos junto con un esquema de votación directa. Para su implementación se utilizará una estructura *kd-tree* que permite realizar búsquedas de los *k*-vecinos más próximos de forma rápida y aproximada en contraposición a la búsqueda exhaustiva.

Más concretamente, en el trabajo realizado, las clases serán los diferentes tipos de documentos. Así, cada clase será modelada utilizando un número de vectores de características locales extraidos de ventanas (o subimágenes) [2] de dimensiones reducidas sobre una o más imágenes de referencia de la clase. De entre todas las ventanas muestreadas, se filtrarán los vectores que representan una clase en el modelo y que a priori no aportan información relevante y se estudiará el tipo de características a utilizar. Posteriormente, las subimágenes que superen el filtro contribuirán al resultado final de la clasificación mediante el sistema de votación directa mencionado.

A los vectores anteriores se les aplicará una transformación basada en el *Análisis de Componentes Principales* (PCA). Como resultado, en el espacio transformado, las componentes de los vectores quedarán ordenadas por varianza y se les aplicará una reducción de dimensionalidad seleccionando aquellas componentes que presenten mayor varianza. El conjunto de vectores resultante (o prototipos) conformará el conjunto de entrenamiento o modelo. Finalmente, los prototipos se insertarán en una estructura *kd-tree* sobre la que se efectuarán las búsquedas de los *k*-vecinos más próximos.

La fase de test consistirá en muestrear la imagen de un documento extrayendo múltiples subimágenes, obteniendo sus vectores de características de la misma forma en que se ha realizado en los vectores de entrenamiento, y clasificándolos. Una imagen de test será asignada a la clase que mayor probabilidad a posteriori presente empleando una regla de clasificación basada en un esquema de votación directa, y rechazada en el caso de no superar un límite mínimo de confianza.

Esta técnica basada en el uso de características locales ha sido aplicada con éxito a otros problemas como el reconocimiento de caras o el de matrículas de coche.

Ante un problema de reconocimiento de formas, es más sencillo plantear estrategias de decisión efectivas en la clasificación cuando las variaciones intraclase son pequeñas y al mismo tiempo las variaciones interclase son grandes. En este sentido, la identificación automática de documentos presenta dificultades propias: por un lado, documentos de la misma clase pueden diferir significativamente respecto de la imagen de referencia ya que pueden contener diferentes cantidades de información cumplimentada; por otra parte, documentos de diferentes tipos pueden llegar a ser muy similares, como por ejemplo, cuando tienen una estructura idéntica o muy similar y difieren sólo en unos pocos caracteres (por ejemplo, plantillas de formularios que difieren sólo en el número de página o en la lengua que están escritos). El uso de características locales puede contribuir a que las prestaciones del sistema planteado sean satisfactorias.

1.3. Objetivos conseguidos y aportaciones

- Creación de una base de datos de documentos obtenidos de distintas fuentes. Gran parte de ellos han sido escaneados y etiquetados a partir de documentos originales en papel.
- Adecuación de las imágenes de referencia mediante selección y boorado manual del contenido cumplimentado.

[2] Se emplearán indistintamente los términos "característica local", "representación local" o "subimagen" para hacer referencia al término ingles *"Local feature"*.

- Estudio bibliográfico del estado del arte en identificación de documentos.
- Implementación de herramientas *software* para la experimentación sobre distintos parámetros relevantes en el proceso de identificación, incluyendo preprocesos digitales, filtros de texturas, extracción de características, esquema de votación directa, búsqueda rápida en *kd-trees*, clasificación mediante *k-vecinos más próximos* y análisis de resultados.
- Adaptación del método de votación directa para la clasificación de documentos basado en la extracción de características locales.
- Realización exhaustiva de experimentos para la optimización de parámetros.
- Estudio de una medida de fiabilidad para el rechazo de documentos clasificados con baja confianza.
- Publicaciones:
  - [Arlandis 11] J. Arlandis, V. Castello-Fos, and J. C. Perez-Cortes, *Filled-in document identification using local features and a direct voting scheme*, In IbPRIA, In press, 2011.

1.4. Plan de la obra

El presente trabajo se ha estructurado en 6 apartados. En esta primera parte de introducción se han señalado las razones y objetivos que han llevado a la realización de esta investigación.

En el capítulo 2 se introduce al lector en los conceptos teóricos básicos dentro del campo del reconocimiento estadístico de formas, centrándose en las técnicas que se emplearán posteriormente en el trabajo.

En el capítulo 3 se hace un repaso al estado del arte dentro del campo de la identificación de documentos como parte específica de la clasificación de documentos. Se muestran las diferentes aportaciones que hasta el día de hoy se han ido realizando y que sirven de punto de partida para este trabajo.

En el capítulo 4 se describe de forma detallada el método empleado. Se explicarán y justificarán todas las decisiones tomadas en el diseño del sistema, tanto en la estructura de éste como en la relevancia dada a los distintos parámetros empleados.

En el capítulo 5 se muestran los resultados de los experimentos realizados, centrándose sobre todo en la variación de los parámetros más significativos que han sido objeto de estudio. Finalmente se mostrarán los resultados obtenidos con la combinación óptima de estos parámetros.

Por último, el capitulo 6 resume las conclusiones alcanzadas tras la realización de la investigación.

Capítulo 2

Conceptos de reconocimiento estadístico de formas

En el presente capítulo, se realiza una introducción al campo del reconocimiento estadístico de formas, centrándose en las aproximaciones relacionadas más directamente con el trabajo de investigación efectuado. A modo de introducción, se presenta un modelo genérico de sistema de reconocimiento de formas. Seguidamente se exponen las bases teóricas que sostienen la aproximación estadística al reconocimiento de formas empleada en este trabajo, haciendo hincapié en el método de características locales y esquema de votación directa tratado en [Paredes 01]. También se describirá brevemente la técnica empleada de mejora de la eficiencia computacional de un clasificador estadístico basada en árboles de búsqueda rápida y aproximada *kd-tree* . El último apartado se ha dedicado a describir la metodología utilizada para estimar el ratio de error del clasificador y una breve introducción al concepto de rechazo.

2.1. Modelo genérico de sistema de reconocimiento de formas inductivo y supervisado

Los sistemas de reconocimiento admiten distintas taxonomías. Desde el punto de vista del modo en que se aborda el problema se pueden considerar dos aproximaciones generales: la aproximación deductiva, que intenta abordar el problema de una forma racional, comprendiendo su naturaleza y buscando la manera de resolverlo a partir de unas ideas lógicas; y la aproximación inductiva, que se usa en los casos en los que el enfoque deductivo no es aplicable, típicamente se utiliza en casos en los que el ser humano es capaz de reconocer fácilmente determinadas formas, pero sin tener un conocimiento lógico de cómo se hace. Este último enfoque será el utilizado el presente trabajo.

La aproximación inductiva lleva asociado el concepto de *aprendizaje*, que puede ser supervisado o no supervisado. El aprendizaje no supervisado se aplica en los casos en los que no se conocen *a priori* las clases en las que está dividido el conjunto de prototipos. En este trabajo se aplicará el aprendizaje supervisado, ya que, como se verá más adelante, se parte de un conjunto de datos previamente etiquetado en un número conocido de clases.

Un sistema de reconocimiento opera funcionalmente en dos modos: *entrenamiento* o *aprendizaje* y *test* o *clasificación*. No obstante, la implementación de un sistema de reconocimiento inductivo supervisado debe pasar por un total de cinco fases bien definidas:

Fase de diseño Análisis del problema. Elección de la metodología, técnicas candidatas y fuentes de

datos.

Fase de entrenamiento Selección de un subconjunto de los objetos disponibles y etiquetado de estos. Implementación del sistema utilizando este modelo al que se llama *conjunto de entrenamiento*.

Fase de validación Selección de otro subconjunto de objetos y etiquetado. Validación del sistema utilizando este conjunto de validación. El objetivo es prever el comportamiento del sistema frente a un conjunto de datos independiente y extraer una estimación del ratio de error asociado.

Fase de test Prueba del sistema utilizando un conjunto de datos independiente de los anteriores (entrenamiento y validación) denominado *conjunto de test*.

Fase operativa Una vez completada la implementación y prueba de la aplicación, el sistema funcionará en modo operativo dentro del entorno al que va destinado.

En la figura 2.1 se muestra un modelo genérico de sistema de reconocimiento de formas donde las diversas tareas consideradas se han agrupado en cinco módulos. Esto no significa que todo proceso de reconocimiento deba pasar por todas y cada una de las etapas descritas, sino que serán las características del problema a resolver las que determinen la combinación final de estas. Los módulos de adquisición y preproceso de la señal son particulares del tipo de señal a tratar (imagen, voz, infrarrojos, etc.) y, por tanto, diferiran sustancialmente en función de su naturaleza. Dado que el presente trabajo está basado en el análisis de imágenes, en adelante, las ideas y conceptos relativos a sistemas de reconocimiento se particularizarán para este dominio. En cuanto a las tareas específicas que conforman cada una de las etapas, estas dependerán del tipo de sistema de reconocimiento. Estas tareas se irán describiendo a lo largo del presente trabajo. No obstante, el cometido y funciones básicas de las etapas de un sistema de reconocimiento de imágenes se describe a continuación.

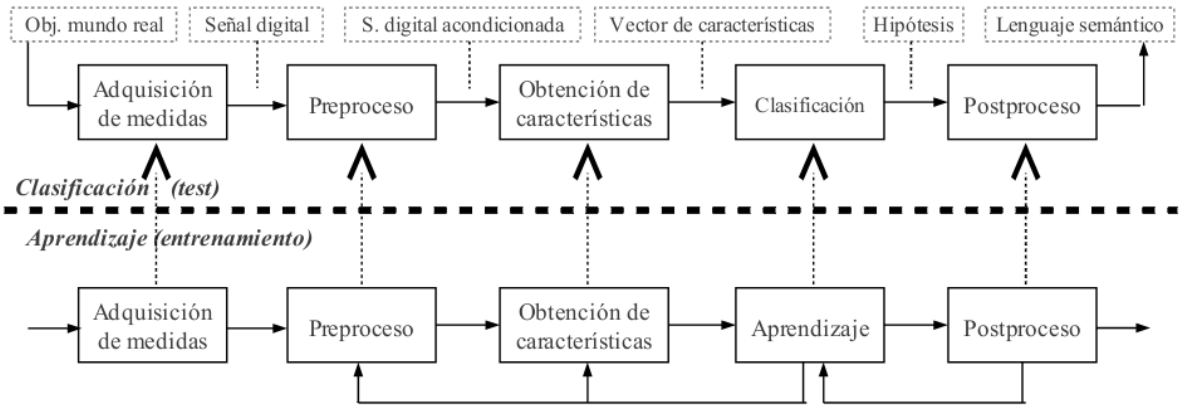


Figura 2.1: Modelo de sistema de reconocimiento de formas.

Módulo de adquisición de datos

Un objeto del mundo real se mide mediante los sensores físicos apropiados, por ejemplo, una cámara de video para la captura de imágenes o un micrófono para registrar la voz. Habitualmente, esta información se plasma en una señal analógica, la cual puede ser preprocesada para mejorar su calidad o

para extraer la parte útil. Una posterior digitalización de la señal permitirá que pueda ser tratada por un ordenador, bien para ser almacenada en un soporte permanente si se trata de prototipos en fase de entrenamiento, bien para disponer de ella de manera inmediata en las fases posteriores. De la señal digital obtenida decimos que pertenece al espacio de representación primario. Este espacio diferirá según la naturaleza de la señal capturada, por ejemplo, una imagen en escala de grises suele representarse como una matriz bidimensional donde cada elemento representa un porcentaje o nivel de gris. La obtención de las medidas de un objeto comprenden los procesos o tareas mostrados en la figura 2.2.

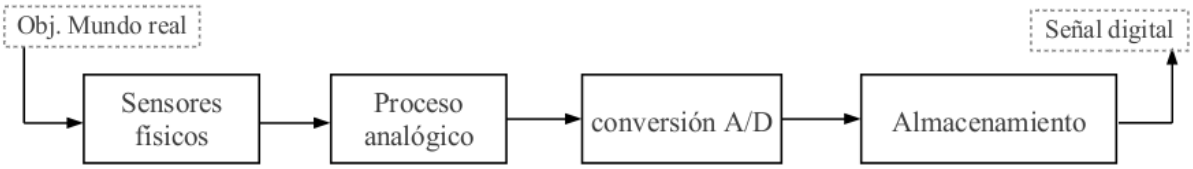


Figura 2.2: Módulo de adquisición de datos.

Módulo de preproceso

El papel del módulo de preproceso, o también denominado, de tratamiento de la señal, es, básicamente, la extracción y acondicionamiento del patrón de interés respecto del resto de la señal. Más concrétamente, tienen cabida la segmentación, eliminación del ruido, normalización y cualquier otra operación que contribuya a definir una representación compacta del patrón. Típicamente, esta nueva representacion sigue perteneciendo al dominio del espacio de representación primario o similar. Ademas, en la fase de entrenamiento se realizará un etiquetado del patrón que permitirá su identificación. El esquema de bloques de la figura 2.3 sintetiza todo esto.

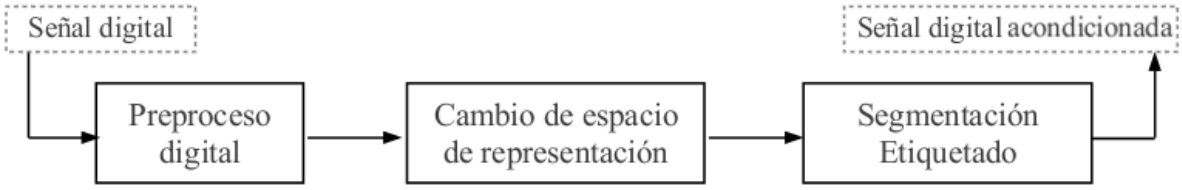


Figura 2.3: Módulo de preproceso.

El tipo de tareas comprendidas en este módulo está fuertemente condicionado por el espacio de repesentación primario de la señal. Así, entre las operaciones más comunes para el tratamiento de imágenes que nos ocupa se encuentran las siguientes:

Escalado y cuantificaciones Ajuste del rango de los datos, reducción de la dimensionalidad.

Filtrado Aplicación de máscaras o filtros de vecindad en el dominio espacial o en el de una transformada. Puede mejorar análisis posteriores de la imagen. Se le suele llamar convolución.

Realzado y modificaciones de histograma Aumento del contraste, retoque de los bordes o zonas tenues en una imagen. Ajustes del rango digital de los datos.

Transformaciones geométricas Cambios de coordenadas, correcciones de perspectiva, escalado de una parte de la imagen, etc.

Operaciones morfológicas Erosión, dilatación y otras.

Eliminación de ruido El ruido puede entorpecer los análisis posteriores. Son fuentes de ruido, tanto el contexto real que envuelve al objeto de interés como el mismo sensor empleado en la adquisición.

Segmentación Extracción del patrón de interés. En escenas donde pueden concurrir múltiples objetos, se puede aplicar una segmentación global, local, adaptativa, etc., para extraer objetos más simples. Por ejemplo, la segmentación de una página en bloques de gráficos y texto, lineas y carácteres.

Etiquetado En fase de entrenamiento se procede a la asignación de un identificador al patrón segmentado paara que pueda ser usado como prototipo. En escenas es necesario etiquetar los distintos objetos o bloques lógicos obtenidos de la segmentación para su reconstrucción posterior.

Módulo de extracción de características

Las características de un objeto son el resultado de la aplicación de ciertas funciones sobre los datos extraidos del objeto (señal digital acondicionada), que ayudan de alguna manera a distinguir la clase a la cual pertenece y que serán usadas como entrada al clasificador. La obtención de características incluye los procesos de obtención, selección y extracción de características (figura 2.4). En este módulo se pretende encontrar aquel conjunto de características que mejor represente al objeto de entrada, esto es, que minimice las diferencias intraclase y maximice las diferencias interclase. En la fase de entrenamiento, es habitual hacer diversas iteraciones sobre este módulo buscando optimizar las prestaciones del sistema. En la fase de test se aplican métodos de extracción y/o selección elegidos durante el entrenamiento de cada uno de los patrones de test. La obtención de características representa la parte menos sistemática del proceso de reconocimiento de formas y depende fuertemente del tipo de tarea en cuestión. Así, la experiencia o la intuición del diseñador del sistema, o la instrospección en el comportamiento humano frente a tipos de tareas similares pueden servir de ayuda en la resolución de esta tarea.

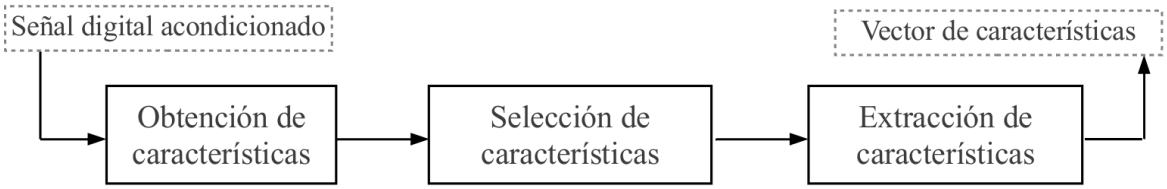


Figura 2.4: Módulo de obtención de características.

Se han considerado tres tareas en el proceso de obtención de características: la obtención que hace referencia a la transformación de la señal digital del dominio primario al dominio considerado más adecuado para tratar el problema, por ejemplo representación vectorial; la selección con la que se eligen aquellas que aportan información valiosa para la identificación del objeto rechazando la información no discriminante o redundante; y la extracción de características que contribuirá a rebajar el coste computacional del proceso tras aplicar un proceso de análisis de componentes principales y la posterior reducción de dimensionalidad. A menudo, los términos extracción y reducción de características son empleados indistintamente. Los criterios de obtención, selección y extracción de características estarán condicionados a los resultados obtenidos en la fase de entrenamiento.

Módulo de aprendizaje

Un clasificador es entrenado tratando de particionar el espacio de características generado por el módulo anterior con el objetivo de minimizar el ratio de error entre los prototipos del conjunto de entrenamiento. La retroalimentación hacia todos los módulos anteriores (figura 2.1) permite al diseñador del sistema optimizar estrategias y elegir de entre las distintas técnicas probadas aquella que mejor rendimiento proporcione de acuerdo con los resultados de validación obtenidos del clasificador.

En algunos casos como el que nos ocupa, es necesaria la combinación de múltiples clasificadores para adaptar la estrategia a la tarea concreta a resolver.

Módulo de clasificación

En fase de test u operativa, el clasificador toma el vector de características del patrón de entrada y lo asigna a la clase que presenta una mayor similitud de acuerdo con el comportamiento inducido durante el aprendizaje. En algúnos casos, la salida del clasificador puede no limitarse a la clase más similar y ofrecer múltiples hipótesis en forma de lista de n-tuplas conteniendo el identificador del objeto, etiqueta de clase (hipótesis) y nivel de fiabilidad asignado a la hipótesis. Frente a un problema por resolver, se puede escoger entre diversos tipos de clasificadores o combinarlos mediante alguna de las técnicas existentes.

Módulo de postproceso

El postproceso que sigue a la clasificación tratará de enmarcar el resultado de la clasificación dentro del contexto semántico al cual pertenecen los objetos. El grado de satisfacción de los resultados obtenidos finalmente servirá para acabar de validar el clasificador o, en algunos casos, puede implicar una vuelta atrás para ajustar o redefinir su salida, particularmente a causa de los niveles de fiabilidad. Se trata de seleccionar la mejor hipótesis para un objeto o para un conjunto de ellos con algún tipo de relación semántica y dar como resultado final del sistema una respuesta que pertenezca a un determinado lenguaje preestablecido.

2.2. La aproximación estadística

Dentro del reconocimiento de formas existen dos aproximaciones: la estructural o sintáctica y la geométrica o estadística. Cada uno de estos dos paradigmas presenta sus ventajas e inconvenientes, de forma que hay aplicaciones típicas en las que la superioridad de una de las dos técnicas es manifiesta. Suele ser la naturaleza de los datos del problema la que determine en gran medida cual de las dos aproximaciones es la más adecuada. Para tratar los problemas de reconocimiento e identificación de documentos se suele emplear habitualmente la aproximación sintáctica, pero en cambio, en este trabajo se tratará de abordar desde el punto de vista de la aproximación estadística.

La aproximación estadística o geométrica está basada en una teoría clásica y muy establecida cómo es la teoría de la decisión. Un objeto es representado en forma de d características o medidas y es tratado como un punto en un espacio vectorial d-dimensional. El significado de las características viene dado, pues, por la posición de los vectores en el espacio (puntos) y las relaciones de distancia entre estos. El objetivo es escoger aquellas características que permiten distribuir los vectores de los patrones de forma que las distintas clases o categorías ocupen regiones compactas y disjuntas. La efectividad del espacio de representación es mayor cuanto más clara sea la separación entre las clases. Así, dado un conjunto de patrones de entrenamiento de cada clase, el objetivo es establecer unas fronteras de decisión al espacio

de características que permitan separar los patrones pertenecientes a las distintas clases. Los métodos estadísticos están basados en la teoría de la decisión estadística, según la cual, las fronteras de decisión son determinadas por las distribuciones de probabilidad de los patrones que pertenecen a cada clase las cuales pueden ser estimadas o aprendidas a partir de metodologías ampliamente conocidas. Probabilidad a priori, probabilidad a posteriori, y fiabilidad de la clasificación son algunos de los conceptos manejados.

Este enfoque es adecuado para resolver problemas donde los datos puedan representarse de forma numérica sencilla y, más específicamente, si su representación en un espacio vectorial euclídeo, métrico o pseudométrico parece natural. Básicamente, la aplicación de técnicas estadísticas consiste a determinar la clasificación de un objeto en una determinada clase o región en función de su posición en el espacio y las distancias que lo separan de otros puntos del espacio. Hay numerosas reglas y técnicas de clasificación que tendrán que resolver como problema destacado la asignación de un objeto situado en regiones fronterizas de dos o más clases llamadas superficies de decisión.

2.2.1. Consideraciones sobre la clasificación estadística

Entrenamiento del clasificador

Cualquier regla de decisión o clasificación tiene que ser entrenada con un conjunto de muestras. Como resultado de esto, el funcionamiento del clasificador dependerá tanto del número de las muestras como de su calidad. Al mismo tiempo, el objetivo de diseñar un sistema de reconocimiento es clasificar futuras muestras de test que se espera sean diferentes a las de entrenamiento. Por lo tanto, optimizar un clasificador para que maximice sus prestaciones ante un conjunto de entrenamiento puede no producir siempre los resultados esperados ante un conjunto de test. La capacidad de generalización de un clasificador se refiere a su funcionamiento sobre patrones de test que no hayan sido empleados en fase de entrenamiento. Cuando a un clasificador se le atribuye una pobre capacidad de generalización suele ser debido a alguno de estos factores: el número de características es demasiado grande respecto del número de muestras de entrenamiento, el número de parámetros no conocidos asociado al clasificador es grande, o bien, el clasificador ha sido optimizado demasiado intensivamente sobre el conjunto de entrenamiento (sobreentrenamiento o *overtraining*); esto es análogo al fenómeno del sobreajuste en regresión cuando hay también un excesivo número de parámetros.

Reducción de la dimensionalidad

Hay dos razones para tratar de conseguir dimensionalidades cuanto más bajas mejor en la representación de objetos: costes y precisión. Un clasificador de puntos de baja dimensionalidad será más rápido y consumirá menos memoria. No obstante, una reducción en el número de características puede implicar una pérdida de poder de discriminación y, por lo tanto, una reducción de la precisión del sistema. Por otro lado, la posibilidad de inclusión de características redundantes o vacías de información discriminatoria a los vectores puede provocar que dos patrones cualquiera puedan acabar siendo similares si son codificados con un número suficientemente grande de este tipo de características.

Es importante distinguir entre selección y extracción de características. El término selección de características se refiere a algoritmos que seleccionan el mejor (supuestamente) subconjunto de características entre el conjunto de características de entrada (previamente obtenido y filtrado por algún dispositivo sensor). Por el contrario, los algoritmos de extracción de características son aquellos que crean nuevas características a partir de transformaciones y/o combinaciones de las características del conjunto original. Así mismo, es habitual que el proceso de extracción de las características preceda el de selección, a pesar de que el conjunto de características original no se vea modificado: en primer lugar, las carac-

terísticas se extraen de la señal digital condicionada proveniente del preproceso y se transforman (por ejemplo, usando análisis de componentes principales o PCA) y después hay una selección para rechazar aquellas que tienen un bajo poder discriminatorio.

Indirectamente, la reducción de la dimensionalidad plantea la necesidad de elección de una función criterio. Un criterio empleado habitualmente es el del error de clasificación asociado a un subconjunto de características, pero el error de clasificación por si mismo puede no ser adecuadamente estimado si el ratio tamaño de muestras a número de características es pequeña. Además de la elección de la función criterio, hay que determinar la dimensionalidad adecuada del espacio reducido, es decir, la *dimensionalidad intrínseca* de los datos. La dimensionalidad intrínseca de los datos determina esencialmente si un conjunto de patrones *d*-dimensional puede ser descrito adecuadamente en un subespacio de dimensionalidad menor de *d*. Por ejemplo, puntos *d*-dimensionales a lo largo de una curva suave tienen una dimensionalidad intrínseca de 1, independientemente del valor de *d*.

Elección de la métrica

En un espacio vectorial donde los puntos se agrupan en regiones propias de una misma clase, el concepto de distancia o medida de disimilitud es importante, puesto que los resultados proporcionados por un clasificador estadístico estarán condicionados por la medida empleada. En la práctica la distancia euclídea es la que se usa normalmente. La métrica euclídea tiene ciertas ventajas, analíticos en particular, puesto que es derivable en todos los puntos. Sin embargo, otras medidas pueden resultar más ventajosas en ciertas ocasiones.

La elección de la métrica que mejor refleje la proximidad de los puntos al espacio no está exenta de dificultades. Cuando las características implicadas son variables multievaluadas, la métrica euclídea requiere el escalado o normalización adecuada de estas. Cuando esto no es fácilmente alcanzable, esta distancia puede resultar inadecuada puesto que las variables con un rango mayor tendrán un peso mayor en la distancia.

Otras métricas que pueden ser usadas son la distancia de Mahalanobis o las distancias de Minkowski, como por ejemplo la distancia L_1 o la L_∞. A pesar de que las métricas son el conjunto de distancias más usadas para variables multievaluadas, otras medidas no métricas como las medidas de disimilitud pueden ser empleadas. Una de estas, el coeficiente de correlación ρ, es también muy usada. Cómo es natural, en cualquier caso, aquellas variables que no tengan relevancia a la hora de discriminar entre clases tendrían que ser excluidas del cálculo de la distancia.

2.2.2. La regla de Bayes

El paradigma de la clasificación ha sido tradicionalmente el marco predominante de los estudios en reconocimiento de formas. En el conjunto de los objetos a reconocer se establece una partición de forma que cada objeto pertenece en una sola clase. Así, ante un objeto no visto, el objetivo consistirá en determinar la clasificación de este, a partir de su representación, en una de las clases posibles .

El proceso de toma de decisión en el reconocimiento de formas estadístico se puede resumir de la manera siguiente: dado un patrón, este tiene que ser asignado (clasificado) en una de las c categorías $w_1, w_2, ..., w_c$ basándose en el vector de d valores de características $x = x_1, x_2, ..., x_d$. Se asume que las características tienen una *masa* o *densidad de probabilidad* (dependiendo de si son discretas o continuas), en forma de función condicionada a la clase, conocida como *densidad condicional de clase*. Así, un vector x de un patrón que pertenezca a la clase w es visto como una observación o punto del espacio d-dimensional extraído aleatoriamente de la función de probabilidad condicional de clase $p(x|w_y)$. Podemos decir que

una regla de decisión da lugar a una partición del espacio de características en c regiones o clases. Las fronteras que separan estas regiones se denominan fronteras de decisión.

Consideramos en primer lugar la posibilidad de que no tengamos información de las características de la muestra a clasificar. La probabilidad de que una muestra cualquiera, de la cual no tenemos ninguna información, pertenezca a la clase w_y es lo que conocemos como *probabilidad a priori* $P(w_y)$. Este valor está asociado a cada clase y no depende de la muestra en concreto. En general, es relativamente fácil y seguro estimar esta probabilidad por simple conteo si disponemos de suficiente cantidad de muestras de identidad conocida extraídas del ámbito natural de una manera uniforme y aleatoria. En el supuesto de que no fuera posible conocer ningún dato del objeto, la decisión más razonable –es decir, aquella que minimizaría el riesgo de error–, sería asignarle la clase con probabilidad a priori más alta:

$$P(w_i) > P(w_j), \quad 1 \leq i, j \leq c\,; \quad i \neq j$$

Esta regla de decisión ciega es, indudablemente, poco útil y supone no hacer uso de ninguna característica de los objetos que pueda contribuir a una clasificación más fiable. La información que las características aportan tiene que emplearse para maximizar la fiabilidad de la clasificación. En este sentido, resultaría muy útil disponer de un mecanismo que nos proporcionara, a partir de estas observaciones, información sobre la *probabilidad a posteriori*, también llamada *probabilidad condicional* $P(w_y|x)$ que un objeto x pertenezca a la clase w_y. La manera de obtenerla es justamente la característica diferenciadora de los distintos métodos de clasificación. Una vez obtenida la probabilidad a posteriori para cada clase, se escogerá aquella clase que presente el mayor valor

$$P(w_i|x) > P(w_j|x), \quad 1 \leq i, j \leq c, \quad i \neq j \quad \Rightarrow \quad x \in w_i \tag{2.1}$$

Este criterio constituye la regla de decisión de Bayes de mínimo error y en ésta se basan de una u otra manera la práctica totalidad de los métodos de clasificación de tipo estadístico.

Cómo hemos visto, el conocimiento de la probabilidad a posteriori es necesario para la aplicación de las reglas de clasificación comentadas. Una buena parte de los métodos existentes hoy por hoy no permiten calcular directamente este valor pero, en cambio, son capaces de estimar a partir del conjunto de entrenamiento las funciones de densidad de probabilidad de cada clase $p(x|w_y)$ en el espacio de representación de las muestras. Con la ayuda de la fórmula de Bayes, podemos obtener la probabilidad a posteriori que la muestra x pertenezca a la clase w_y si conocemos, para esta clase, el valor de la función de densidad en este punto:

$$P(w_i|x) = \frac{p(x|w_i)P(w_i)}{p(x)},$$

y, ya que el denominador no depende de la clase, la regla de decisión de Bayes se puede reescribir como:

$$p(x|w_i)P(w_i) > p(x|w_j)P(w_j), \quad 1 \leq i, j \leq c, \quad i \neq j \quad \Rightarrow \quad x \in w_i$$

Como se ha dicho anteriormente, las posibilidades a priori son conceptualmente fáciles de estimar, no así las funciones de densidad. Se distingue entre aquellos procedimientos que no hacen uso de ningún conocimiento respecto de la naturaleza de las funciones de densidad a estimar y aquellos que requieren asumir que estas responden a una determinada distribución especificada paramétricamente. Los primeros son llamados métodos no paramétricos y los últimos paramétricos. Hay además un conjunto de técnicas conocidas como funciones discriminantes en las cuales se asume una forma paramétrica, no para las funciones de probabilidad, sino para las superficies de decisión.

2.3. Esquema de votación directa basado en los *k*-vecinos más próximos

En los sistemas tradicionales de clasificación basados en el paradigma estadístico, cada clase se representa por un vector de características, y se aplica una regla de discriminación para clasificar un vector de test representado también por un vector de características. En este trabajo, se utilizará una técnica de clasificación basada en características locales, es decir cada documento (tanto de entrenamiento como de test) estará representado por un conjunto de vectores de características locales. Cada uno de estos vectores puede clasificarse dentro de una clase diferente, y será un sistema de votación directa el que decidirá finalmente la clase asignada al documento de test.

La técnica de extracción de características consiste en representar un documento con un conjunto de regiones de la imagen. Cada una de estas regiones está codificada en forma de vector, al cual se le exigirá que cumpla unas determinadas condiciones para que sea realmente representativo utilizando distintos filtros (varianza, entropía, etc.). Una región de la imagen de tamaño $v \times w$ estará representada por un vector de $v \times w$ dimensiones, cada una de las cuales contiene el valor de intensidad de un píxel de la región. Con objeto de reducir el coste computacional se aplicará una reducción de dimensionalidad a e dimensiones utilizando el método PCA (*Principal Component Analysis*).

La extracción de características se puede formalizar de la manera siguiente: Sea $I = \{I_1, \ldots, I_n\}$ un conjunto de entrenamiento de n imágenes que representan d clases diferentes. Para cada imagen I_i, se obtienen m_i vectores de características que son proyectados con el método PCA a un espacio de e dimensiones. Sea $X_i = \{x_{i1}, \ldots, x_{im_i}\}$, el conjunto de vectores e-dimensionales asociado con la imagen I_i, y sea $T = \bigcup_{i=1}^{n} X_i$ el conjunto global de vectores. Cada vector x_i llevará asociada una etiqueta de clase $\omega^i \in \{\omega_1, \ldots, \omega_d\}$ que es la etiqueta de clase de la imagen I_i. Obviamente, todas las imágenes pertenecientes a la misma clase de documentos tendrán la misma etiqueta.

El procedimiento de clasificación utilizado está estrechamente relacionado con una familia de técnicas referidas de forma habitual como *"sistemas de votación directa"* [Mohr 97]. De hecho, está basado en la aplicación de la conocida regla de los *k*-NN (*k* vecinos más próximos) al conjunto de vectores que representan una imagen de prueba, utilizando el vector global de conjunto T como conjunto de referencia o conjunto de prototipos. Más formalmente, podemos presentar la técnica de clasificación en el marco estadístico de las "combinaciones de clasificadores" [Kittler 98].

Sea Y una imagen de test. Siguiendo el marco probabilístico convencional, Y puede ser perfectamente clasificada en la clase $\hat{w}$, que tiene la máxima probabilidad a posteriori:

$$\hat{w} = \arg\max_{1 \leq j \leq d} P(\omega_j \mid Y) \tag{2.2}$$

Aplicando el proceso de extracción basado en características locales descrito anteriormente, la imagen Y estará representada por un conjunto de vectores $m_Y = \{y_1, \ldots, y_{m_Y}\}$. Se puede ver el clasificador 2.2 como una combinación de m_Y clasificadores, uno para cada vector de características de Y. Asumiendo la independencia entre los vectores y_i, se puede escribir $P(\omega_j \mid Y)$ como el producto de las probabilidades a posteriori asociadas a cada vector de características, y 2.2 se convierte en:

$$\hat{w} = \arg\max_{1 \leq j \leq d} \prod_{i=1}^{m_Y} P(\omega_j \mid y_i) = \arg\max_{1 \leq j \leq d} \sum_{i=1}^{m_Y} \log P(\omega_j \mid y_i) \tag{2.3}$$

Esto es conocido comunmente como la *"regla del producto"* para combinaciones de clasificadores. Un inconveniente importante de esta regla es que las probabilidades muy pequeñas tienden a dominar el

resultado de la combinación, provocando pobres estimaciones de $P(\omega_j \mid Y)$ y un bajo rendimiento de la clasificación. Para aliviar estos problemas se suele utilizar la llamada *"regla de la suma"* [Paredes 01]. Esta regla se puede ver como una forma de suavizar el efecto de las probabilidades pequeñas.

En las situaciones reales, la mayoría de las probabilidades a posteriori $P(\omega_j \mid Y)$ suelen tener valores cercanos a 1 o cercanos a 0. Aquellos cercanos a 1 se pueden aproximar linealmente como:

$$\log P(\omega_j \mid y_i) \approx P(\omega_j \mid y_i) - 1$$

Obviamente, para aquellas probabilidades cercanas a 0, esta aproximación lineal no es muy adecuada, produciendo valores significativamente mayores que los correctos. Sin embargo, el error introducido realmente produce un efecto de suavizado beneficioso que tiende a compensar la pobre estimación de dichas pequeñas probabilidades. Con todo esto, utilizando la aproximación lineal, la ecuación 2.2 se puede reescribir como:

$$\hat{w} = \underset{1 \leq j \leq d}{\arg\max} \sum_{i=1}^{m_Y} P(\omega_j \mid y_i) \tag{2.4}$$

que corresponde a la mencionada *"regla de la suma"* para combinaciones de clasificadores.

Cabe destacar que los vectores de características que provienen de subimágenes con contenido cumplimentado, o con cualquier otro tipo de ruido, introducen al clasificador vectores "ruidosos" que serán generalmente estimados con bajas probabilidades. La *"regla de la suma"* produce un efecto de suavizado de estas bajas probabilidades. Es más, los vectores "ruidosos" realmente tendrán un efecto beneficioso ya que los vectores mal clasificados tendrán tendencia a distribuirse entre diferentes clases.

En el caso que nos ocupa, las probabilidades a posteriori se estiman directamente con la regla de los *k*-vecinos más próximos. Sea k_{ij} el número de vecinos de y_i pertenecientes a la clase ω_j. Asumiendo que el número medio de vectores que representan a todas las imágenes de entrenamiento de cada clase es más o menos constante, una buena estimación de $P(\omega_j \mid y_i)$ es:

$$\hat{P}(\omega_j \mid y_i) = \frac{k_{ij}}{k},$$

y, usando esta estimación en 2.4, nuestra regla de clasificación se convierte en:

$$\hat{w} = \underset{1 \leq j \leq d}{\arg\max} \sum_{i=1}^{m_y} k_{ij} \tag{2.5}$$

Esto es, se selecciona la clase $\hat{w}$ con el mayor número de *"votos"* acumulados sobre todos los vectores pertenecientes a la imagen de test. De esta forma se justifica el por qué estas técnicas son conocidas como *"sistemas de votación"*.

En [Paredes 01] se describe un clasificador similar muy utilizado para biometría facial a partir de un sistema de votación directa basado en *k-NN*.

2.4. Búsqueda rápida en *kd-trees*

El *kd-tree* es un árbol clásico entre los que se utilizan para la búsqueda del vecino más cercano. El nombre *kd-tree* viene de *k-dimensional tree*, en el que la k representa la dimensión de los datos del espacio de representación. El *kd-tree* es un árbol binario que contiene en cada nodo intermedio información acerca de una coordenada que divide en dos el conjunto de datos del subárbol correspondiente al nodo, y en las hojas contiene puñados (*buckets*) de prototipos. Durante la fase de clasificación se recorre el árbol siguiendo un esquema de ramificación y poda para encontrar el vecino más cercano. Tanto en el preproceso (construcción del árbol) como en la clasificación propiamente dicha se utilizan las coordenadas de los prototipos, por lo que esta estructura de datos y el algoritmo de búsqueda necesitan un espacio de representación con vectores (espacio de vectores). Aunque se han desarrollado muchas mejoras y optimizaciones, es el algoritmo de referencia cuando se utilizan distancias euclídeas.

Construcción del *kd-tree*

La idea principal del algoritmo para la construcción del *kd-tree* a partir de un conjunto de prototipos P es la siguiente: encontrar un hiperplano que divida el conjunto P en dos subconjuntos y proceder recursivamente con los subconjuntos. El principal aspecto a resolver es la elección del hiperplano y de la coordenada que va a servir para dirigir la búsqueda a un lado u otro del hiperplano, la coordenada discriminante.

Para intentar conseguir que cualquier prototipo tenga la misma probabilidad de estar a un lado o a otro del hiperplano y por tanto que el árbol resulte lo más equilibrado posible, se suele elegir el hiperplano de forma que se sitúe en la mediana de los valores de la coordenada discriminante. Además, la coordenada discriminante debe ser aquella que tenga una mayor amplitud, es decir, aquella para la que la diferencia entre la coordenada mínima y máxima sea la mayor en valor absoluto.

El árbol se construye de la siguiente forma: en cada nodo, que representa un conjunto de prototipos (el nodo raíz representa a todo el conjunto de entrenamiento), se elige la coordenada discriminante y se obtiene la mediana de los valores de dicha coordenada en los prototipos del conjunto; a continuación, se divide dicho conjunto en dos subconjuntos utilizando la mediana, situando en un subconjunto aquellos prototipos para los que el valor de la coordenada discriminante sea menor o igual que el de la mediana, y en el otro subconjunto los prototipos cuya coordenada discriminante sea mayor que la mediana. A continuación, se crean recursivamente los árboles asociados a cada uno de los subconjuntos.

El proceso termina cuando el tamaño del conjunto de prototipos es menor o igual que el valor fijado como tamaño de un puñado, y en este caso el nodo sería una hoja.

Búsqueda en el *kd-tree*

El proceso de búsqueda en el *kd-tree* es recursivo: dada una muestra desconocida x, en un nodo cualquiera del árbol (que no sea una hoja) se compara la coordenada de x que es discriminante para ese nodo (c) con el valor de corte v (la mediana de las coordenadas discriminantes), y se procede en la dirección más cercana según esa coordenada. Si $x[c] + d_{nn} \leq v$ (donde d_{nn} es la distancia al vecino más cercano hasta el momento), el hijo derecho de ese nodo no puede contener al vecino más cercano y por tanto no es necesario buscarlo en ese nodo; de forma similar, si $x[c] - d_{nn} \geq v$ el hijo izquierdo no es necesario visitarlo. Si el nodo es una hoja, la muestra se compara con todos los prototipos contenidos en ella.

2.5. Recuperación de documentos

Uno campos de estudio con gran actividad en la actualidad es el de la recuperación de información o *document retrieval*. Este auge surge con las nuevas tecnologías y la necesidad de analizar y medir las prestaciones de los sistemas de almacenamiento masivo de documentos [Díaz 03].

La recuperación de información es el proceso que permite obtener, de un fondo documental, los documentos adecuados a una determinada demanda de información por parte de un usuario. Este proceso engloba el conjunto de acciones referidas a la identificación, selección y acceso a los recursos de información necesarios para resolver el problema del usuario.

Cuando se produce una necesidad informativa, mediante una estrategia de búsqueda más o menos complicada, interrogamos al conjunto de documentos, con el fin de obtener una respuesta que satisfaga la demanda. Para saber en qué medida la respuesta es satisfactoria, es necesario evaluar los resultados. Desde este punto de vista, la evaluación es la etapa final de la creación de un sistema.

La importancia de la evaluación en recuperación de información está muy ligada a la fase de investigación ya que sin unas medidas eficaces y estandarizadas, y colecciones experimentales adecuadas para este fin, no se pueden hacer evaluaciones, ni lo que es más importante, no se pueden comparar los sistemas de un modo fiable.

Los documentos pueden ser recuperados o rechazados al establecer la comparación entre la pregunta y la base de datos. El conjunto de documentos recuperados se divide, salvo en los sistemas perfectos, en dos grupos: documentos relevantes recuperados, es decir aquellos que se han recuperados correctamente y los no relevantes, recuperados erróneamente que provocan ruido en la salida. Los documentos no recuperados, que a su vez se dividen en los relevantes, rechazados por el sistema de manera errónea y los no relevantes, rechazados de manera correcta por el sistema. Esto mismo lo podemos ver en la figura 2.5.

Para encontrar un paralelismo entre la recuperación de información y la identificación de documentos tratada en este trabajo es necesario introducir el concepto de rechazo en la identificación de documentos. Un clasificador debe ser capaz de detectar documentos de clases para las que no ha sido entrenado y rechazarlos. Con este concepto, la tarea de clasificación puede ser vista como recuperación de información: los documentos de clases conocidas (relevantes desde el punto de vista de recuperación de información) deberán ser clasificados (recuperados) correctamente y el resto de documentos de clases desconocidas (no relevantes) deben ser rechazados.

Desde este punto de vista, es posible utilizar las medidas estandarizadas ampliamente utilizadas en el campo de recuperación de información para evaluar los resultados de un clasificador. A continuación se detallan algunas de estas medidas.

2.5.1. Precisión y exhaustividad

La precisión es la proporción de material recuperado realmente relevante, del total de los documentos recuperados. Es conocida también como factor de pertinencia o ratio de aceptación. El resultado de esta operación está entre 0 y 1. Así, la recuperación perfecta es en la que únicamente se recuperan los documentos relevantes y por lo tanto tiene un valor de 1.

Esta medida está relacionada con dos conceptos, el de ruido y el de silencio informativo. De este modo, cuanto más se acerque el valor de la precisión a 0, mayor será el número de documentos recuperados que no le sirvan al usuario y por lo tanto el ruido que encontrará será mayor.

La fórmula del ratio de precisión es:

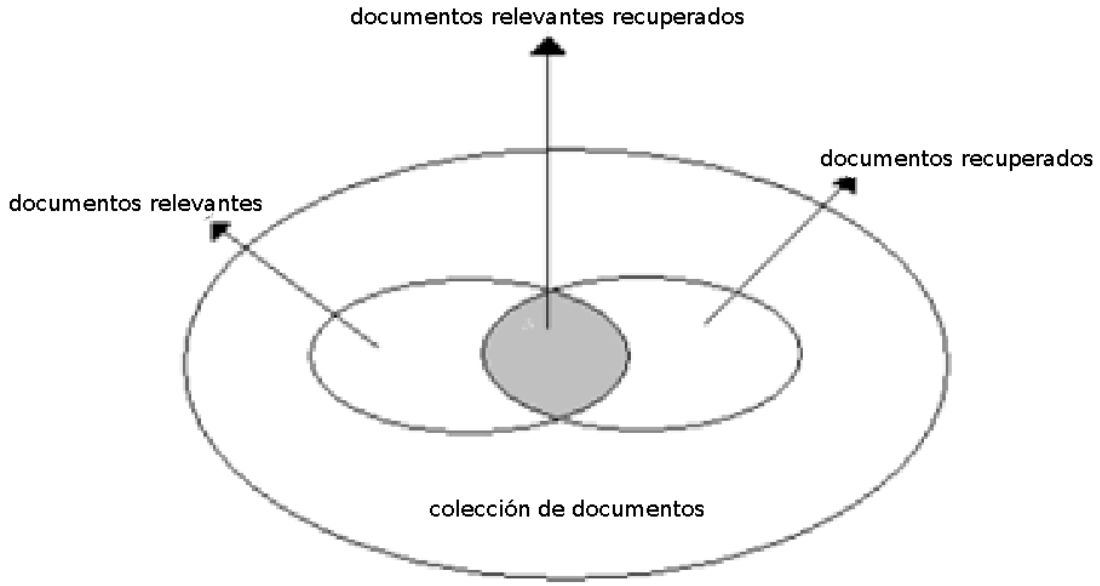


Figura 2.5: Esquema de recuperación de documentos.

$$precision = \frac{documentos\ relevantes\ recuperados}{documentos\ recuperados}$$

La exhaustividad es el otro concepto más utilizado en la evaluación de los sistemas de recuperación. Muchos autores, por influencia del término inglés la denominan *"recall"* o *"rellamada"*. Es la proporción de material relevante recuperado, del total de los documentos que son relevantes en la base de datos, independientemente de que éstos, se recuperen o no.

La exhaustividad es inversamente proporcional a la precisión y se calcula de la siguiente manera:

$$exhaustividad = \frac{documentos\ relevantes\ recuperados}{documentos\ relevantes}$$

Si el resultado de este cálculo tiene como valor 1, tendremos la exhaustividad máxima, ya que hemos encontrado todo lo relevante que había en la base de datos, por lo tanto no tendremos ni ruido ni silencio informativo: la recuperación será perfecta.

Para utilizar estos términos en la identificación de documentos será necesario definir un índice de fiabilidad en la clasificación que permita ordenar los documentos de más a menos fiables. El índice de fiabilidad es una forma de cuantificar la relevancia de los documentos y debe representar el nivel de seguridad que tiene un clasificador de que cierto documento pertenezca a la clase en que ha sido identificado.

Una medida ampliamente utilizada que aporta un buen indicador de la calidad de un clasificador es el índice de exhaustividad para un determinado valor de precisión, *"recall at x % precision"*, que indica el ratio de exhaustividad que se obtiene con un índice de fiabilidad predeterminado que permite un x % de precisión. Habitualmente se utiliza el *"recall at 100 % precision"*.

Capítulo 3

Estado del arte en identificación de documentos

En la clasificación de documentos, tradicionalmente, se ha dedicado una significativa cantidad de esfuerzo a desarrollar aproximaciones basadas en agrupar documentos con un cierto grado de similitud semántica como pertenecientes a una misma clase o categoría. Sin embargo, en algunas aplicaciones, como las relacionadas con la digitalización y extracción de datos, entre otras, las clases deben ser definidas para representar tipos particulares de documentos. En este caso, la tarea es comunmente conocida como "identificación de documentos", y los métodos de agrupamiento o *clustering* no son adecuados. En la mayor parte de estas aplicaciones, la identificación de la imagen de un documento es requerida en primera instancia, antes de cualquier otro proceso específico.

Se han utilizado muchos tipos de características para la clasificación de imágenes de documentos. Están relacionados con el diseño de los documentos, primitivas de texturas, reconocimiento de caracteres y cadenas, códigos de silueta, detección de marcos, *visual salient features*, transformaciones globales y proyecciones de imágenes, o la detección de las estructuras semánticas de los bloques.

En el ámbito de recuperación de la información, cuando no existe contenido cumplimentado, es decir, contenido que puede y suele variar entre diferentes documentos de la misma clase , la identificación de documentos puede ser visto como una tarea de detección de duplicados [Doermann 98]. En este caso, los planteamientos tienen que hacer frente a las diferencias entre las instancias de documentos, como la resolución, sesgo, distorsiones y calidad de imagen, velocidad y robustez, así como, al manejo de bases de datos muy grandes.

La mayoría de trabajos que tratan con documentos cumplimentados se restringen a la identificación de formularios. Muchos de ellos se basan en el análisis de estructuras globales y locales [Fan 01], [Ohtera 04], [Mandal 05]. Las características estructurales se limitan generalmente a documentos que contienen marcos, celdas, lineas, bloques o elementos similares, y no son de ayuda cuando diferentes tipos de documentos tienen estructuras similares. Otros trabajos se basan en el uso de códigos de carácter o de cadena para lograr identificar los documentos [Sako 03], así como, en computar las densidades de píxeles dentro de algunas regiones de la imagen [Heroux 98]. Dentro de los documentos tipo formulario existen aplicaciones específicas encaminadas a la identificación de cupones [Nagasaki 06], recibos bancarios [Ogata 03] o documentos administrativos, por ejemplo [Ting 96].

El propósito de este trabajo es hacer frente a la tarea de clasificación de documentos con total independencia de los diseños, distribuciones, tamaños y cantidad de contenido rellenado de una manera eficiente. Por lo tanto, la aproximaciónes anteriores pueden no ser apropiadas, bien porque utilizan características globales, se centran en tipos de documentos específicos, o no son capaces de manejar

documentos con contenido cumplimentado.

Así, son escasos los trabajos encontrados en la literatura referidos a la identificación de imágenes de documentos cumplimentados. Algunos trabajos más directamente relacionados con este proyecto, son los presentados por Parker [Parker 10] and Sarkar [Sarkar 06], [Sarkar 10]. Sarkar [Sarkar 06] presenta una metodología para seleccionar y clasificar puntos de anclaje o *"anchor points"* a partir de imágenes de documentos. La selección de *anchor points* está basada en el uso de características destacadas rectangulares de Viola&Jones en el canal de luminancia. Para cada clase de documento, se obtiene la distribución de probabilidad de la lista de características locales (incluyendo las coordenadas globales de posición) mediante un modelo de Independencia Condicional Latente (LCI). Se clasifica una imagen emparejando su lista de características con modelos generativos específicos de la clase por el criterio de máxima verosimilitud, y se asigna a la clase cuya distribución empírica está más cercana, de acuerdo con el valor de KL-divergencia de Kullback-Leibler. Esta correspondencia es bien conocida en la comunidad de clasificación y recuperación de textos, donde las observaciones son listas de palabras de longitud variable. Recientemente, Sarkar [Sarkar 10] propone una metodología completa para seleccionar *anchor points* basados en subimágenes elegidas de forma aleatoria y aplicando sucesivos refinamientos expandiendo y ordenando los puntos candidatos utilizando dos medidas de calidad.

Parker [Parker 10] propone y compara tres métodos para la selección de *anchor points*. El primero está basado en dos criterios: la "acción gráfica" y la minimización de la distancia intra-clase. Los otros dos métodos intentan la selección de *anchor points* que maximicen la función de KL-divergencia, una medida de separación de dos distribuciones de probabilidad; uno de las distancias entre *anchor points* dentro de una muestra dada de una clase de documento, y el otro de las distancias de estos *anchor points* a los documentos de distintas clases. Parker afirma que el rendimiento del sistema propuesto de identificación de formularios puede ser estimado de manera teórica mediante la KL-divergencia. El autor muestra los resultados de los experimentos de los tres métodos utilizando una base de datos personalizada de formularios extraídos del IRS, donde sólo un tipo de documento contenía datos rellenados y se utilizaron diez formularios para entrenar el sistema. La principal conclusión de los experimentos es que el uso de la información inter-clase para seleccionar los puntos de anclaje de una clase mejora el rendimiento del sistema (estimado mediante la KL-divergencia). Este método implica el uso de varios documentos de cada clase para entrenar el sistema, y puede ser necesario efectuar un elevado número de operaciones de correlación para que la selección de *anchor points* sea robusta frente a las traslaciones de la imagen, algo necesario en la fase de producción.

Capítulo 4

Preproceso, extracción y selección de características

La primera tarea a realizar en un trabajo de identificación de documentos es, obviamente, la digitalización de estos mediante un escáner. A partir de este momento, se dispone de una imagen digital para cada documento escaneado. Típicamente, la imagen de un documento se compone de un fondo de píxeles blancos y píxeles negros en primer plano, aunque se pueden encontrar otras combinaciones, como documentos en escala de grises, color o combinaciones heterogéneas de fondos y primeros planos. El primer plano de un documento se compone sobre todo de texto (en muchos casos con diferentes aspectos como los tipos de letra, estilos de escritura, letras mayúsculas, texto en negrita, diferentes tamaños, etc), aunque otros objetos como imágenes, gráficos, logotipos, o marcos son también frecuentes. Por lo general, las áreas de texto también incluyen patrones de fondo, y un patrón de fondo también pueden estar presente en la mayoría de la superficie de un documento.

En este capítulo, se realizará un recorrido por los principales puntos en los que se ha basado este proyecto. Se justificará el empleo de varias funciones de preproceso de imágenes con el objeto de mejorar el rendimiento del sistema, tanto en velocidad de computación como en tasa de aciertos y se describen las técnicas empleadas para la extracción de características de las imágenes.

Así, se realizará un submuestreo particularizado, que permite reducir el número de características a extraer de cada imagen disminuyendo el tiempo de proceso en la fase de test; los filtros de varianza y diferencia de entropía, con los que la selección de características locales es más efectiva ya que se descartan aquellas que aportan menos información a la clasificación; la reducción de dimensionalidad, que permite el algoritmo PCA proyectando los vectores de características en un espacio más disccriminativo que el original y permitiendo prescindir de un número de dimensiones sin perdida de información discriminativa; o el uso de las coordenadas de posición de cada ventana como características globales del documento.

La secuencia de operaciones que se aplicarán a las imágenes de documentos antes del proceso de clasificación es la siguiente:

1. Conversión a PGM
2. Corrección de la rotación
3. Normalizado a tamaño equivalente a un A4 escaneado a 300 dpi
4. Suavizado y umbralizado de las imágenes

5. Escalado
6. Extracción de características con submuestreo
7. Filtro de diferencia de entropía
8. Filtro de varianza
9. Analisis de componentes principales
10. Incorporación de caracteristicas globales a los vectores

En las secciones siguientes se describen detalladamente cada uno de estas operaciones.

4.1. Preproceso de imágenes

Los documentos del corpus presentan varios inconvenientes que deben ser resueltos antes de realizar los experimentos. Como se verá en el punto 5.1, las imágenes de los documentos tienen distintos formatos (PNM, TIFF), y dentro de estos formatos existen diferencias en la representación de los píxeles (1-bit y 8-bits). El primer paso del preproceso de las imágenes ha sido convertir todos los documentos a un mismo formato. Utilizando la herramienta *convert* del paquete *ImageMagick* se han convertido todos los documentos a archivos de tipo PGM o *Portable GrayMap*. En el caso de documentos binarios, se ha cambiado el rango de valores para que ocupen 8-bits y puedan ser tratados en las mismas condiciones que el resto de documentos.

Este primer paso resulta trivial, pero todavía existen efectos no deseables que conviene corregir en los documentos, como la rotación o las distintas tonalidades de gris dentro de documentos de una misma clase debido a defectos en el proceso de escaneado. Estas correcciones requieren de determinados preprocesos más complejos que se detallan en los siguientes puntos.

4.1.1. Corrección de la rotación

El problema a la hora de corregir posibles rotaciones de un documento es el desconocimiento del ángulo al que ha sido rotado. En algunos casos, el documento es escaneado perfectamente, y en otros, sufre desviaciones aleatorias en los dos sentidos y de magnitudes diferentes. El problema se reduce a detectar el ángulo de rotación del documento [Andreu-Cerezo 10].

Así, durante una primera fase se calculará el ángulo de rotación de la imagen. Para ello se aplicará inicialmente el operador Sobel para la detección de bordes, eliminando de la imagen aquella parte de los objetos que no es de interés. El operador Sobel se emplea en el procesamiento de imágenes, utilizándose frecuentemente en algoritmos de deteccióon de bordes. Cuando es aplicado a una imagen en escala de grises, calcula el gradiente de la intensidad de brillo de cada píxel, dando la dirección del mayor incremento posible. El resultado recoge el cambio en cada píxel analizado y la orientación a la que tiende ese borde. En la figura 4.1 se puede apreciar el resultado de aplicar Sobel sobre una imagen original.

Se ha utilizado la Transformada de Hough para calcular el ángulo de rotación. La Transformada de Hough es una herramienta que permite detectar curvas partiendo de su expresión analítica, siendo muy frecuente su utilización en la localización de rectas, circunferencias y elipses.

La Transformada de Hough requiere de una importante suma de operaciones de cálculo, que dependen de la cantidad de píxeles negros de la imagen, por ello, antes de aplicarla, se realiza un filtrado a la

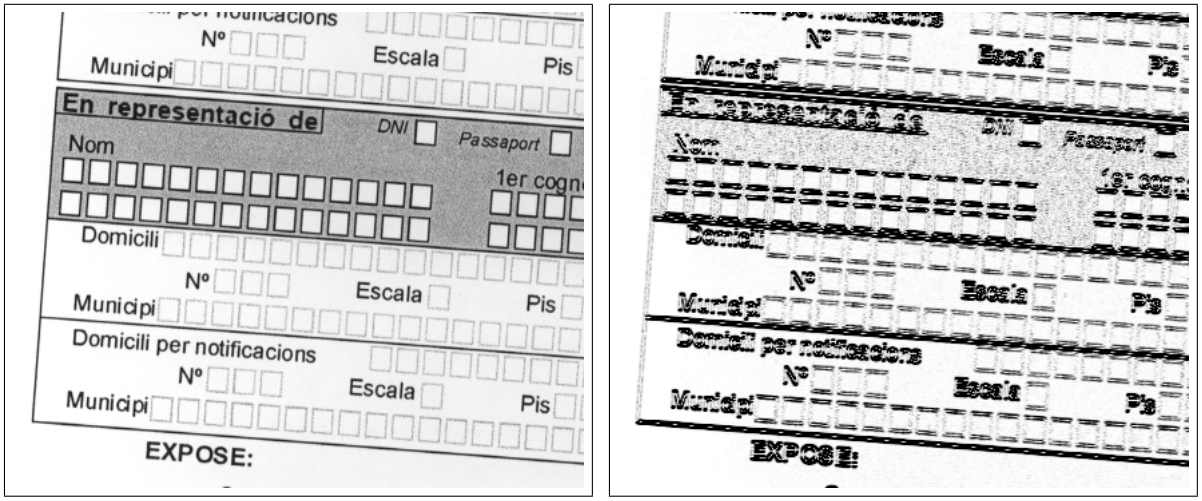


Figura 4.1: Imagen antes y después de aplicar Sobel.

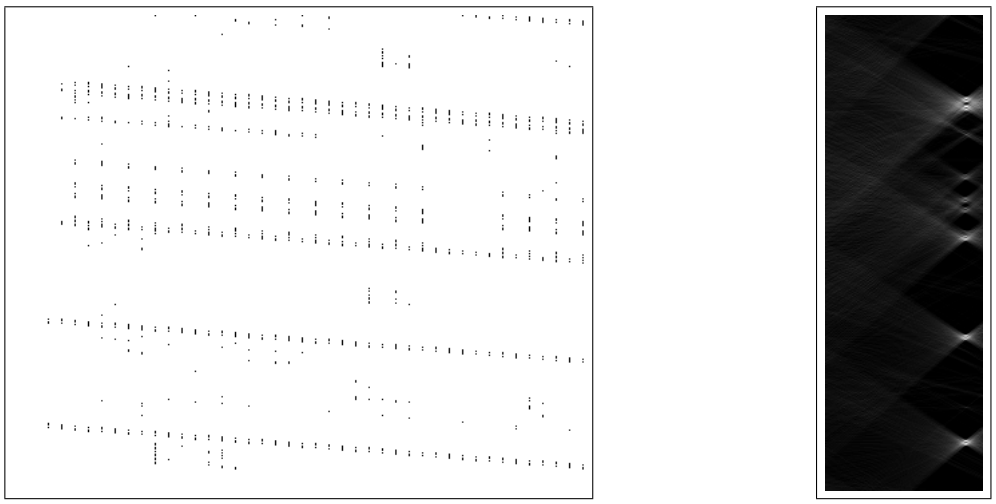


Figura 4.2: Imagen filtrada y mapa de Hough resultante.

imagen y, a su vez, se reduce la búsqueda de rectas a aquellas que tengan entre -5 y 5 grados respecto del eje ortogonal horizontal, reduciéndose considerablemente el número de cálculos a realizar. El filtrado consiste en buscar pequeños segmentos horizontales (10 píxeles), estos se transformarán en un píxel en la imagen de entrada a Hough, como se puede apreciar en la figura 4.2.

Partiendo de la ecuación de la recta en su forma explícita, $y = a * x + b$, donde a es la pendiente y b la ordenada en el origen, si (x_i, y_i) es un punto de la recta (a, b), entonces $y_i = a * x_i + b$. El objetivo es pues encontrar los puntos de la imagen que satisfacen la ecuación de la recta para los distintos valores de a y b. Dada una recta, los parámetros a y b son constantes, siendo (x, y) las variables. Partiendo de la imagen lo que conocemos son varios puntos que pertenecen a la recta, por lo que despejando de la ecuación la pendiente de la recta, nos quedará de la siguiente manera, $b = -a * x_i + y_i$, donde (x_i, y_i) son las coordenadas de un píxel dado. Si se representa esta ecuación para los valores a y b se obtiene una nueva recta donde (a, b) son ahora las variables. Esta transformación entre el plano imagen (coordenadas x-y) y el espacio de parámetros (coordenadas a-b) se denomina Transformada de Hough.

En este espacio de parámetros, cada punto (x, y) de la imagen se convierte en una recta de pendiente

Figura 4.3: Imagen Resultado

$-x$ y ordenada en el origen y, que representa el lugar geométrico de todas las rectas que en el plano imagen pasan por el punto (x, y). Por tanto, la intersección en el espacio de parámetros de dos rectas (x_i, y_i) y (x_j, y_j) da lugar al par (a', b') que corresponde a la recta que, en el plano imagen, pasa por los puntos (x_i, y_i) y (x_j, y_j). De este modo, los puntos que forman una recta en el plano imagen se transforman en el espacio de parámetros en rectas que interseccionan en un mismo punto (a', b').

Como se ha visto, las rectas del espacio de parámetros se obtienen a partir de los píxeles en negro de la imagen, dichas rectas forman el llamado Mapa de Hough, como se puede ver en la figura 4.2. Los puntos donde más se acumulan intersecciones, dentro del espacio de parámetros, producen las rectas más significativas en el espacio imagen, de las que se puede obtener con facilidad el ángulo α predominante de las mismas y con el cual se puede compensar la rotación.

$$\alpha = arctan(\frac{y_j - y_i}{x_j - x_i})$$

Finalmente, una vez obtenido el ángulo de rotación, se compensa la rotación de la imagen. (Figura 4.3).

4.1.2. Normalizado del tamaño de las imágenes

El tamaño de las imágenes se ha normalizado para que ocupen una superficie similar a un A4 escaneado a 300 dpi, conservando el *aspect ratio* del documento original.

4.1.3. Filtros globales

Las imágenes digitalizadas no presentan siempre una calidad adecuada para ser usadas por un clasificador. Dos escanéres distintos, por ejemplo, pueden producir imágenes muy diferentes al digitalizar el mismo documento, bien porque las imágenes obtenidas presentan distintas tonalidades de gris, bien porque añaden ruido a la imagen, etc. Lo mismo sucede si un documento es digitalizado a formatos de imagen distintos (escala de gris, color o b/n). Es habitual que los documentos presenten este tipo de problemas, por lo que se hace necesario un tratamiento previo de las imágenes para tratar de mini-

mizar estas diferencias, de lo contrario, el método de digitalización podría influir negativamente en la clasificación de determinados tipos de documento.

Para solucionar estos problemas, a cada imagen de documento se le ha aplicado un proceso global de filtrado basado en dos etapas:

1. **Filtro de suavizado:** cada píxel de la imagen se sustituye por el valor medio de los píxeles de una matriz de convolución de 3x3 en la que el centro de la matriz es el píxel original. De esta forma se consiguen varias mejoras:
   - transformar los tramados de imágenes escaneadas originalmente en b/n a tonalidades de gris
   - suavizar los perfiles de todas las zonas de interés
   - eliminar parte del ruido originado en el proceso de escaneo
2. **Filtro de umbralizado o binarización de la imagen:** cada píxel de la imagen se convierte en blanco o negro dependiendo de si el valor original del pixel es mayor o menor que un valor dado (umbral). En este caso se ha utilizado un umbral del 70 % del valor máximo de un píxel, o lo que es lo mismo, un valor de $255 \times 0{,}7 \simeq 178$. Por tanto, los píxeles con valor superior a 178 pasarán a tener un valor de 255 (blanco), y el resto se convierten en píxeles de valor 0 (negro).

En la figura 4.4 se muestran algunos ejemplos de documentos originales y sus correspondientes imágenes tras el proceso de filtrado.

4.1.4. Escalado

Con el escalado de las imágenes de documentos no se pretende corregir ningún defecto sino acelerar el tiempo de proceso de cada documento.

Colateralmente, el escalado ejerce un efecto de suavizado del contenido del documento, igualando los bordes de zonas idénticas que en la escala original podían aparecer con formas distintas tras el escaneado. Una vez aplicado un factor de escala, todos los documentos deben tener la misma superficie.

4.2. Extracción de características

Para poder aplicar el algoritmo de clasificación de los *k*-vecinos más próximos, es necesario utilizar una representación vectorial o modelo matemático de los documentos. En el conjunto de prototipos del modelo, cada clase estará representada por una serie de vectores de características locales potencialmente discriminativos respecto del resto de clases. Para la clasificación de un nuevo documento, se extraerán sus características locales en forma de vectores y se compararán con los vectores del conjunto de entrenamiento, clasificándose en la clase más parecida de acuerdo a las reglas de los *k*-vecinos descritas en el capítulo 2.

Cada vector de características locales se forma a partir de los píxeles de una determinada región de la imagen. Por ejemplo, se puede formar un vector de 25 dimensiones a partir de una ventana de la imagen de 5 x 5 píxeles sin más que concatenar los valores de cada píxel, empezando por la esquina superior izquierda de la matriz y recorriéndola de izquierda a derecha, y de arriba a abajo. Para poder emplear distancias euclideas entre estos vectores (u otros cálculos como la reducción PCA) será necesario que todos ellos tengan la misma dimensión. Por lo que en cada experimento, las subimágenes correspondientes a cada característica local deben tener el mismo tamaño.

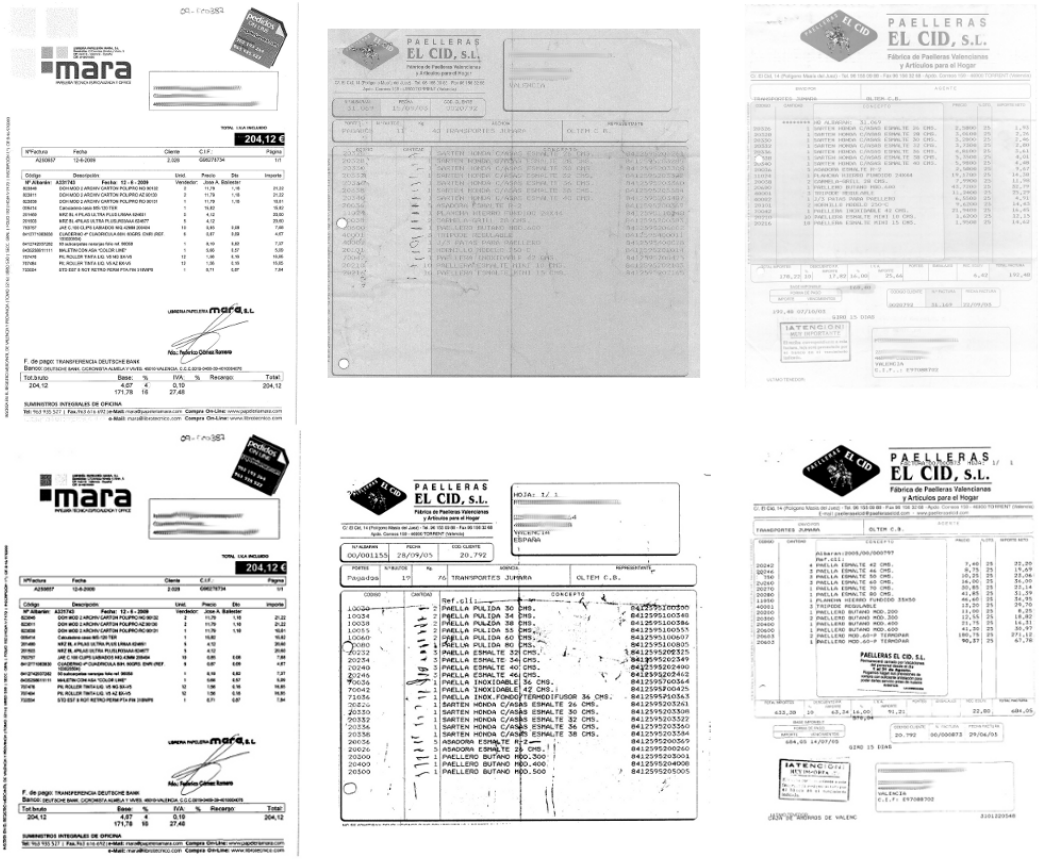


Figura 4.4: Ejemplos de documentos tras aplicación de filtros globales. Imágenes originales en la fila superior e imágenes filtradas en la fila inferior.

Resulta evidente que una buena elección de los criterios de selección de las características locales será fundamental para el correcto funcionamiento y las tasas de acierto del sistema.

En los siguientes apartados se describen detalladamente los criterios seguidos en este trabajo para la selección y extracción de características locales. Algunos de estos criterios persiguen una mejora en la tasa de acierto y fiabilidad de los resultados, mientras que otros buscan mejorar el tiempo de proceso requerido en el proceso de clasificación.

4.2.1. Selección de características locales

El criterio elegido para la selección de características locales se basa en la aplicación de una combinación de filtros. En una primera etapa se aplicará un filtro de textura con el objeto de eliminar subimágenes con texturas demasiado comunes en ciertos tipos de documentos y que, por tanto, no son lo suficientemente discriminativas. En la segunda etapa se tratará de eliminar las características que, habiendo superado el primer filtro, no aportan suficiente información. Para ello se eliminarán las subimágenes con bajo índice de varianza.

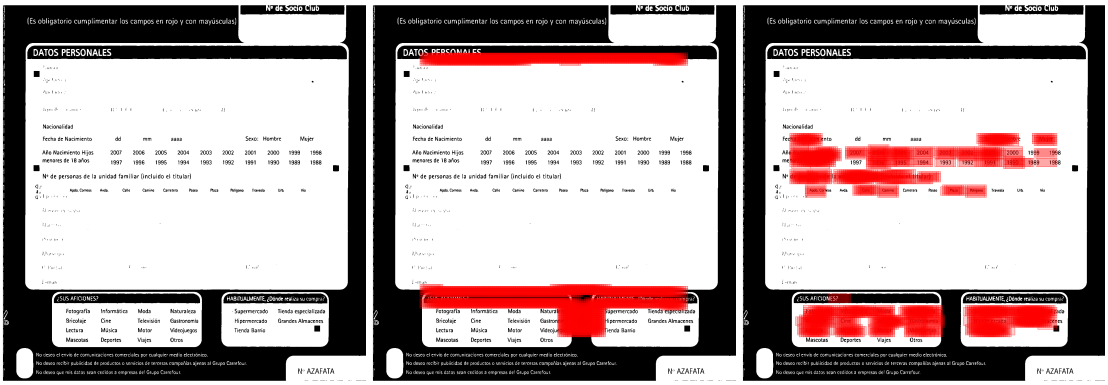


Figura 4.5: Efecto del filtro de diferencia de entropía en las características seleccionadas. De izquierda a derecha: imagen original, características seleccionadas sin filtrado, características seleccionadas tras aplicar el filtro de diferencia de entropía.

Filtros de textura

En el punto 4.2.1 se verá como el criterio final de selección de características se basa en la varianza de éstas. Pero antes, es importante asegurarse de que las subimágenes de alta varianza son suficientemente discriminativas, por lo que en un paso previo se aplica un filtrado de texturas con el objeto de eliminar aquellas subimágenes que, aún teniendo alta varianza, son demasiado comunes en distintos tipos de documentos y, por tanto, no aportan información discriminativa al proceso de clasificación. Por ejemplo, documentos con una considerable cantidad de marcos de alto contraste (cuadros blancos sobre fondo negro, por ejemplo), o defectos bastante comunes de escáneres que producen franjas verticales de color oscuro, pueden provocar una fuerte caída en los resultados obtenidos por el clasificador. Si estas zonas no son discriminantes respecto de otros tipos de documentos el clasificador fallará de forma estrepitosa.

Para conseguir que este tipo de características no sean seleccionadas, en favor de otras características de alta varianza y mejor capacidad de discriminación, se han utilizado los filtros de textura basados en la matriz de co-ocurrencia de las imágenes.

Haralick propuso un conjunto de 14 medidas de textura basadas en la dependencia espacial de los distintos tonos de gris [Haralick 73]. En este trabajo se ha experimentado con todas ellas tratando de identificar los valores que presentan las subimágenes a filtrar en cada una de estas medidas. Una vez obtenidos estos valores, el filtrado de este tipo de caracteristicas es tan fácil como eliminar las subimágenes que presenten valores similares a los experimentados.

Las 14 medidas propuestas por Haralick son: segundo momento angular, contraste, correlación, varianza, inversa de la diferencia de momentos, suma de la media, suma de la varianza, suma de la entropía, entropía, diferencia de la varianza, diferencia de la entropía, medida de correlación (I y II) y máximo coeficiente de correlación.

Cada una de estas propiedades se pueden aplicar a pares de píxeles desplazados a una distancia fija. Variando distancia y angulo entre dos píxeles se pueden obtener medidas características de distintos tipos de textura. En el caso que nos ocupa, se han aplicado los filtros considerando únicamente pixeles adyacentes, es decir, se han obtenido las matrices de co-ocurrencia aplicadas a píxeles vecinos en los 4 angulos posibles (0°, 45°, 90°y 135°).

Después de diversos análisis se ha determinado un filtrado basado en la medida de diferencia de entropía es suficiente para descartar las texturas no deseadas. El cálculo de la diferencia de entropía se realiza de la siguiente forma:

$$f_{11} = -\sum_{i=0}^{N_g-1} P_{x-y}(i) \log P_{x-y}(i)$$

siendo

$$P_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i,j)$$

con

- $k = \{0, 1, \ldots, N_g - 1\}$
- $\mid i - j \mid = k$
- $P(i,j)$ es la probabilidad (o frecuencia) de la entrada (i,j)-ésima de la matriz de co-ocurrencia
- N_g es el número de posibles tonos de gris de la imagen

Filtro de varianza

Los formularios y documentos administrativos son normalmente documentos con gran parte de su superficie de color blanco. Al extraer las características locales de estos documentos es importante que éstas contengan información relevante, y por tanto, conviene descartar las características que no aportan información específica del documento. En particular, conviene descartar los vectores que se forman a partir de regiones uniformes (todos los píxeles de la misma intensidad o muy similar).

Para ello, se puede establecer un filtro basado en la varianza de los vectores, descartando aquellos que no superen un determinado umbral, o seleccionando los n mejores de una lista ordenada por varianza.

La varianza de un vector se define como la varianza estadística de sus componentes, es decir, dado un vector de n dimensiones $\vec{x} = (x_1, ..., x_n)$ su varianza σ^2 viene determinada por la fórmula

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

donde μ es la media aritmética de las componentes del vector $\vec{x}$.

El método empleado en este trabajo consistirá en combinar las dos propuestas mencionadas anteriormente, es decir, exigir a todos los vectores de características extraídos un valor mínimo de varianza σ^2_{min}, y entre ellos, elegir sólamente los n mejores (si hay suficientes), descartando el resto como representantes del documento.

En los experimentos realizados se tratará de ajustar los dos valores planteados σ^2_{min} y n con el propósito de conseguir la mejor tasa de acierto posible.

Cabe destacar un par de observaciones acerca del filtrado de catacterísticas locales:

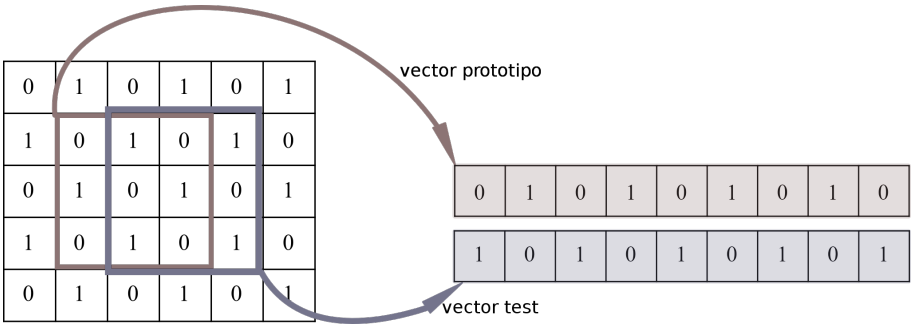


Figura 4.6: Ejemplo de la problemática del submuestreo. Dos ventanas de un mismo documento extraídas en las fases de entrenamiento y test no coinciden en ninguna componente del vector pese a ocupar parcialmente la misma zona de la imagen.

- Dado que a las imágenes de entrenamiento se les ha eliminado la información cumplimentada, los vectores seleccionados contendrán siempre información de la parte estática del documento. En cambio, las imágenes de documentos de test no han sido limpiadas, de modo que la selección puede incluir muchos vectores de la parte cumplimentada del documento. Para compensar este problema, en los documentos de test se extraerá un número de características superior al extraído en los documentos de entrenamiento.
- Hay que tener cuidado con las ventanas de baja varianza. Si no se establecen bien los filtros, y alguna clase de documentos contiene vectores de baja varianza en sus prototipos, pueden actuar como sumidero en la clasificación de cualquier documento de test que también incluya este tipo de características. Esto provocaría una disminución considerable en la tasa de acierto del clasificador.

4.2.2. Submuestreo y orla de vecindad

Una estrategia que puede ser aplicada para mejorar los tiempos de proceso es el *subsampling* o submuestreo, aplicable tanto para la obtención de los prototipos como para la obtención de los vectores de test. En un procedimiento sin submuestreo se extraen todas las características locales asociadas a cada píxel del documento. Esto conlleva tratar con un número muy elevado de características y gran cantidad de información redundante. Para evitar este derroche de recursos se puede establecer una distancia de separación entre características locales conocida como paso de submuestreo, de forma que los vectores seleccionados aporten información al proceso de clasificación. Es posible establecer pasos de submuestreo distintos en los dos ejes del documento. Debe ser tenida en cuenta la relación entre el paso de submuestreo y la forma o tamaño de la ventana de características locales.

Esta estrategia tiene un serio inconveniente cuando no existe solapamiento entre los vectores de entrenamiento y test. Si un documento de test es escaneado con un pequeño desplazamiento inferior al paso de submuestreo respecto del documento de referencia utilizado en el entrenamiento, las ventanas extraidas en los dos documentos jamás coincidirán y la clasificación fallará.

En la figura 4.6 se muestra un ejemplo extremo de lo que puede suceder con dos vectores casi idénticos tras un desplazamiento de un píxel.

Para evitar el problema de "no solapamiento" de prototipos y vectores de test se ha empleado la siguiente estrategia: una vez extraidas las características de los documentos de entrenamiento y reducido el conjunto de estas tras aplicar los filtros de diferencia de entropía (4.2.1) y de varianza (4.2.1) se añaden al conjunto los vectores de características asociados a los píxeles vecinos de cada una de ellas. A este conjunto de vecinos se le llamará "orla de vecindad".

De esta forma se garantiza que si un vector extraido en la fase de test es relevante para la clasificación de un documento, su correspondiente prototipo en el documento habrá sido incluido en el conjunto de entrenamiento y la clasificación correcta será mucho más probable. Asimismo, el tamaño de la ventana de vecinos debe guardar una relación directa con el tamaño de submuestreo empleado en el procedimiento.

Al incluir los vectores de la orla de vecindad en el conjunto de datos de entrenamiento se empeora ligeramente el tiempo de entrenamiento del sistema. Por contra, el coste computacional de la fase de test disminuye drásticamente con el empleo del submuestreo, además, no se ve afectado de manera significativa por el aumento del número de prototipos ya que la búsqueda en kd-tree tiene un coste logarítmico.

4.2.3. Análisis de componentes principales

El análisis de componentes principales (PCA) pretende reducir la dimensionalidad del espacio de representación de los objetos a partir de proyecciones lineales. Las proyecciones son elegidas de forma que la varianza total de los objetos en la nueva representación sea máxima. Es un método no supervisado, ya que no se tiene en cuenta la clase de los objetos.

La transformación PCA consiste en lo siguiente:

Sean N objetos representados por los vectores de características $\{x_1, x_2, ..., x_N\}$ tales que $x_i \in \mathbb{R}^n$. Se considera una transformación lineal que transforme el espacio original n-dimensional en un espacio m-dimensional, con $m < n$. Las nuevas representaciones y_i se calculan como sigue:

$$y_i = W^T x_i \qquad i = 1, 2, ..., N$$

siendo las m columnas de $W \in \mathrm{R}^{n \times m}$ ortonormales.

Sea S_c la matriz de covarianzas de los vectores de características originales definida como:

$$S_c = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T$$

donde $\mu \in R^n$ representa la media de los vectores de características originales. S_c tiene dimensión $n \times n$ y es simétrica, por tanto, tendrá n vectores propios $\{\phi_1, \phi_2, \dots \phi_n\}$ y n valores propios $\{\lambda_1, \lambda_2, \dots \lambda_n\}$ que cumplen la relación:

$$S_c \Phi = \Phi \Lambda \tag{4.1}$$

donde,

Φ es la matriz de vectores propios $\phi_1, \phi_2, \dots \phi_n$

Λ es la matriz de valores propios $\begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}$

Los ϕ_i son ortogonales entre si, por tanto, la matriz Φ define una transformación lineal ortogonal, que define una nueva base y no deforma, por tanto, el espacio original:

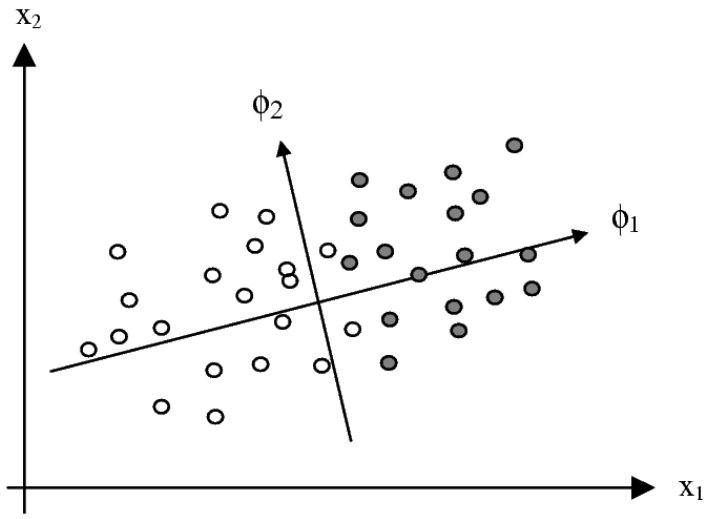


Figura 4.7: Ejemplo de transformación PCA en un espacio de dos dimensiones. El vector propio ϕ_1 es el que mayor varianza explica. Una reducción del espacio original definido por los ejes $x1, x2$ en el espacio definido por el eje ϕ_1 mantiene una clara separación de las observaciones entre las clases.

$$y_i = \Phi^T x_i \tag{4.2}$$

después de esta transformación, los vectores y_i tienen como matriz de covarianzas:

$$\Phi^T S_c \Phi \tag{4.3}$$

dado que Φ es ortogonal ($\Phi^T\Phi = I$) y usando (4.1), la nueva matriz de covariazas es Λ. Con esta transformación se consigue decorrelar los valores de las nuevas dimensiones (que serán las nuevas características). Además, cada nueva dimensión, generada por un vector propio, tendrá como varianza el valor propio correspondiente.

Si se define $W = [\phi_1, \phi 2, \ldots \phi_m]$, siendo $\phi_1, \phi_2, \ldots \phi_m$ los m vectores propios de mayor valor propio, se habrá escogido el subespacio lineal de m dimensiones que más cantidad de la varianza original de los datos explica. A este proceso se le llama reducción de la dimensionalidad.

Un ejemplo de la aplicación de esta transformación se puede ver en la figura 4.7.

La reducción PCA permite mejorar el tiempo de proceso del sistema al reducir el número de cálculos necesarios en la etapa de clasificación. En los experimentos realizados se probarán proyecciones en distinto número de ejes con el objetivo de encontrar el mínimo número de dimensiones necesarias para conseguir resultados óptimos.

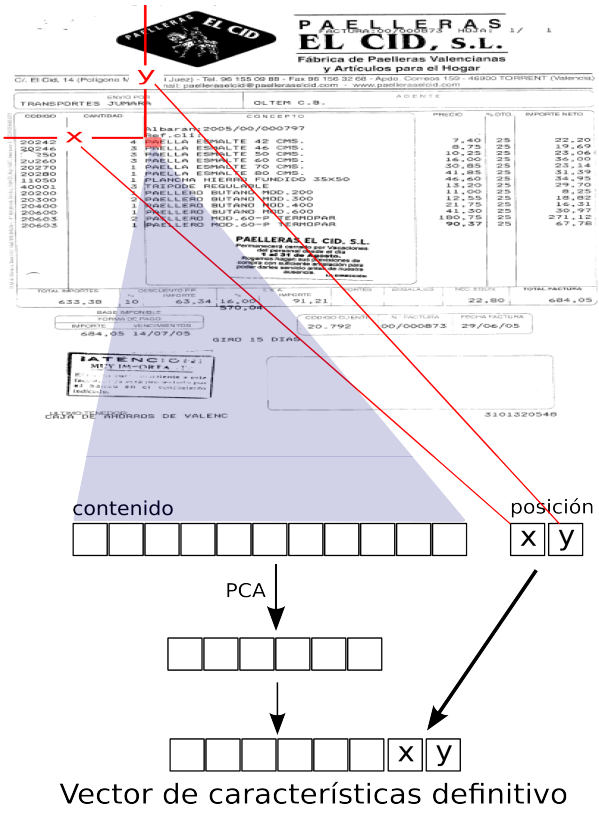


Figura 4.8: Composición final de un vector de características

4.2.4. Coordenadas de posición de las ventanas de características locales

En este punto, tras aplicar los pasos anteriores, cada característica extraida está representada por un vector de n dimensiones que contiene información sobre el contenido de una determinada región del documento sin importar la localización de esta región. Esto puede suponer un problema porque documentos distintos pueden contener gran cantidad de subimágenes similares, por ejemplo, porciones de texto impreso en un mismo tipo de fuente. Se podría mejorar la información aportada por un vector si éste incluyera la localización de la ventana que contiene dichas subimágenes dentro del documento. Esto es tan simple como incorporar las coordenadas de la ventana como dos dimensiones más del vector de características.

Con esta operación se obtiene un vector final de características multimodal con componentes de distinta naturaleza. Para evitar este problema se ha procedido a normalizar las dos componentes de localización de la ventana respecto de la primera coordenada PCA.

También es importante evaluar la influencia o peso de las coordenadas de posición dentro del cálculo de la distancia euclídea entre vectores. Por esto, estas coordenadas no se tendrán en cuenta durante la transformación PCA, y de esta forma se podrán aplicar distintos pesos a estos valores (después de haber sido normalizados), observando cómo afectan al resultado final.

En la figura 4.8 se describe de forma gráfica este proceso.

Capítulo 5

Experimentos

5.1. Descripción del corpus

Para la realización del trabajo ha sido necesario construir una base de datos de documentos aptos para el tipo de estudio que se pretende. La primera intención era utilizar un corpus estándar con el que se pudieran comparar los resultados, pero no se ha encontrado ninguno que se adapte completamente a las necesidades de este proyecto. El único disponible en la literatura es el corpus NIST SD6 [Dimmick 92], pero sólo contiene 20 clases de documentos, por lo que se ha decidido completarlo con documentos provenientes de otros orígenes. Finalmente, el corpus empleado está compuesto por:

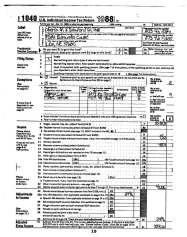
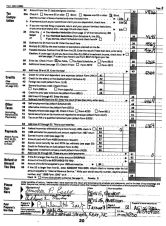
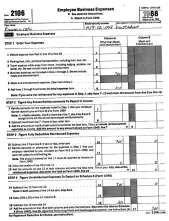
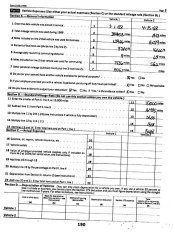
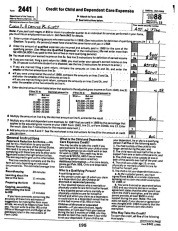
- 20 clases de la base de datos estándar SD6 NIST. Todos elllos son formularios de tasas del gobierno de Estados Unidos cumplimentados a mano.
- 47 clases de documentos escaneados por el autor del proyecto a partir de facturas, recibos bancarios, etc. Todos ellos con contenido impreso o manuscrito variable y con distintos tamaños y proporciones.

En todos los casos, los documentos han sido digitalizados a partir de originales impresos en papel mediante un procedimiento mecánico de escaneado, por lo que algunos presentan defectos de rotación. Dependiendo del origen, algunos documentos están escaneados en escala de grises, otros en blanco y negro. En el corpus hay documentos de distintos formatos, tamaños y tipos, por lo que será necesario un preproceso previo para tratar de transformarlos a un único formato.

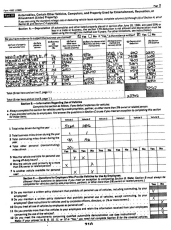
La tabla 5.1 muestra una descripción detallada de las clases de documentos utilizada en los experimentos, indicando las etiquetas asignadas a cada clase, una miniatura de una muestra de cada clase, el tamaño aproximado en píxeles de cada imagen, el tipo de imagen (formato de archivo y profundidad del píxel), el tipo de documento (formulario, factura, albarán o recibo), el origen (atendiendo a los tres orígenes detallados anteriormente) y el número de muestras disponibles para la experimentación.

En los experimentos con opción de rechazo se han utilizado un conjunto de 200 documentos, compuesto por formularios y documentos administrativos de distintos orígenes y formatos pertenecientes a 200 clases, todas ellas distintas entre sí y, por supuesto, distintas a las 67 clases del *corpus principal*.

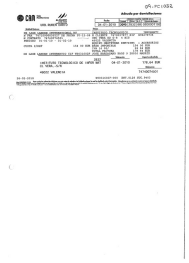
Cuadro 5.1: Corpus utilizado en el proyecto.

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
1040		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
1041		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
2106		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
2107		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
2441		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**

Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
4562		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
4563		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
6251		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
AGAG		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**11**
ALDA		1648x719	TIFF 1-bit B/N	Recibo	IDF1	**13**
AME_		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**

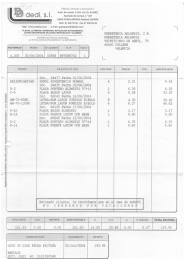
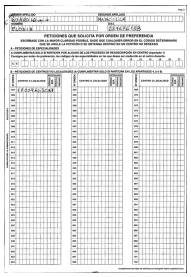
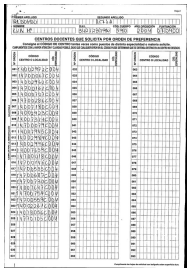
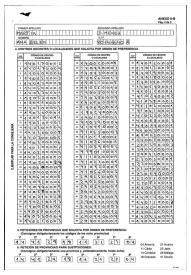
Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
ARCO		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**
BCD_		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**13**
BNCJ		2362x1181	TIFF 8-bit Gris	Recibo	IDF1	**13**
CAM_		1656x2339	PNM 1-bit B/N	Recibo	IDF1	**11**
CENS		3508x2480	TIFF 8-bit Gris	Formulario	IDF1	**13**
CIDA		2480x2362	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**

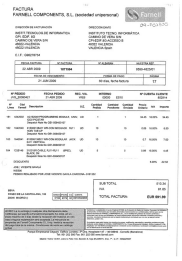
Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
CIDF		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**
COLT		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**9**
COMU		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**10**
CORR		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**13**
CRDT		2480x1181	TIFF 8-bit Gris	Recibo	IDF1	**13**
DEDA		2480x1771	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**

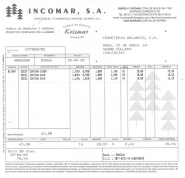
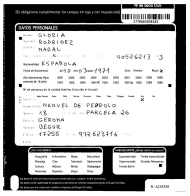
Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
DEDF		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**
EDU1		2410x3549	PNM 8-bit Gris	Formulario	IDF1	**13**
EDU4		2438x3537	PNM 8-bit Gris	Formulario	IDF1	**13**
EDU7		2413x3513	PNM 8-bit Gris	Formulario	IDF1	**13**
EHLS		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**

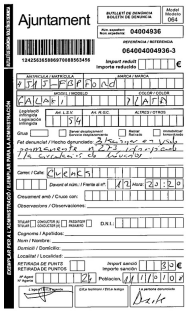
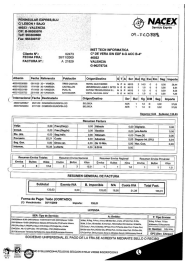
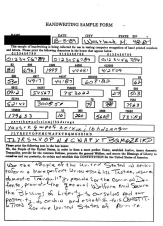
Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
EMVI		2480x3508	TIFF 1-bit B/N	Formulario	IDF1	**13**
FARN		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**13**
FOXN		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**13**
IBDL		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**
IBER		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**13**

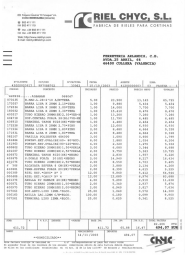
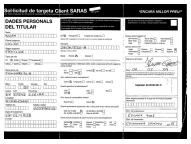
Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
INCO		2480x2362	TIFF 8-bit Gris	Fact/Alb	IDF1	**12**
LEDE		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**
LOTE		1530x1600	TIFF 1-bit B/N	Formulario	IDF1	**13**
MARA		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**12**
MONX		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**

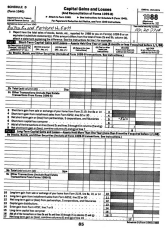
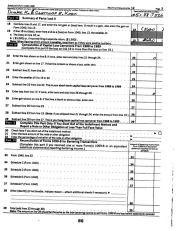
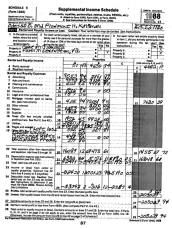
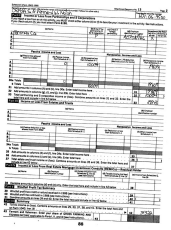
Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
MULT		794x1382	TIFF 1-bit B/N	Formulario	IDF1	**13**
MURC		2480x1175	TIFF 1-bit B/N	Recibo	IDF1	**13**
NACX		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**11**
NIST		2560x3300	PNM 1-bit B/N	Formulario	IDF1	**13**
ONO_		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**13**

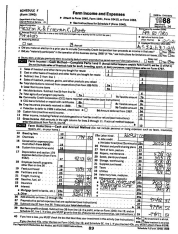
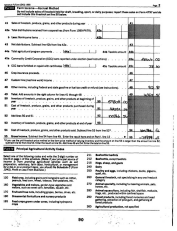
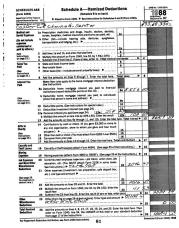
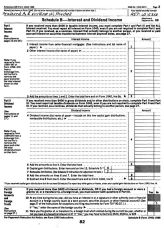
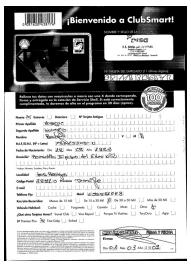
Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
RICO		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**13**
RIEL		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**
ROLS		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**
SARA		3406x2517	TIFF 1-bit B/N	Formulario	IDF1	**13**
SCC1		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**

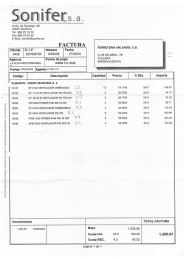
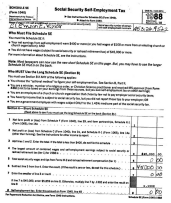
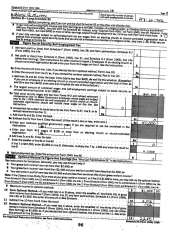
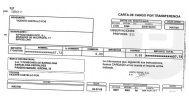
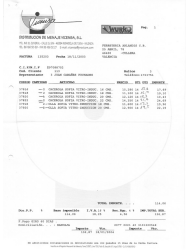
Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
SCC2		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
SCD1		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
SCD2		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
SCE1		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
SCE2		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**

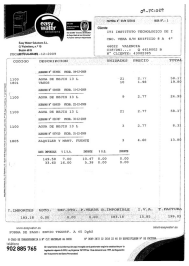
Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
SCF1		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
SCF2		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
SCHA		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
SCHB		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
SHEL		2542x3551	TIFF 1-bit B/N	Formulario	IDF1	**13**

Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
SONI		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**
SSE1		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
SSE2		2560x3300	PNM 1-bit B/N	Formulario	SD6 NIST	**13**
TIMB		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**
UNOE		2480x1181	TIFF 8-bit Gris	Recibo	IDF1	**13**
VICE		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**

Cuadro 5.1 – Continuación

Etiqueta	Miniatura	Tamaño	Tipo	Documento	Origen	nº muestras
WATR		1656x2339	PNM 1-bit B/N	Fact/Alb	IDF1	**9**
YOI2		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**
YOIG		2480x3507	TIFF 8-bit Gris	Fact/Alb	IDF1	**13**

5.2. Preproceso

En los experimentos descritos en esta sección y las siguientes se ha utilizado una imagen de referencia por cada clase para el conjunto de entrenamiento, el resto de imágenes se ha utilizado para el conjunto de test. Las imágenes de entrenamiento se han revisado, y en los casos en los que ha sido necesario, se han limpiado para eliminar el contenido cumplimentado.

Se han aplicado varias técnicas de preproceso a toda la base de datos de documentos. En primer lugar, se ha aplicado una corrección de la rotación a todos los documentos. Se ha aplicado un filtro de suavizado utilizando una matriz de convolución de 5×5 y se ha aplicado un proceso de binarizado con un umbral del 70 %. Para evitar los inconvenientes originados por las distintas resoluciones de adquisición, así como los distintos formatos de imagen, cada imagen se ha normalizado en tamaño

para que ocupe la misma superfície en número total de píxeles (equivalente a la de un documento A4 escaneado a 300dpi) preservando su relación de aspecto original.

En el paso de binarizado, los píxeles de cada imagen han sido convertidos a negro (0) o blanco (255), es decir, se han eliminado los valores de gris. Posteriormente, la fase de normalizado a tamaño A4 ha devuelto algunos de los píxeles a valores intermedios debido al efecto de suavizado que produce este proceso. Éste es el motivo por el que no se han utilizado imagenes binarias con píxeles de un solo bit, ya que al final del proceso, habrá valores de gris dentro del rango [0,255].

La última fase de preproceso ha consistido en escalar los documentos a distintos tamaños y observar la evolución del error obtenido. La figura 5.1 muestra el error de clasificación para distintos valores de escalado de los documentos. Se han aplicado diversos factores de escala entre $1/4$ y $1/12$ (partiendo de las imagenes de area A4 a 300dpi). En cada caso el tamaño de la ventana de características locales se ha escalado aplicando el mismo factor.

5.3. Optimización de parámetros

Las fases de selección y extracción de características requieren del ajuste de distintos parámetros. Para ver como afectan todos ellos al rendimiento del sistema, se han realizado pruebas exhaustivas tratando de probar gran cantidad de combinaciones. En estas pruebas se han combinado los siguientes parámetros:

- *Escalado del documento*. Con el objetivo de reducir el coste computacional, se han aplicado a los documentos distintos factores de escala entre $1/4$ y $1/12$ (partiendo de las imagenes de area A4 a 300dpi). En cada caso el tamaño de la ventana de características locales se ha escalado aplicando el mismo factor.
- *Tamaño de ventana*. Se ha experimentado con varios tamaños de ventana, que incluyen potencialmente subimágenes con distinto número de caracteres de texto u otros objetos. Las ventanas de 80 píxeles de ancho y 30 píxeles (respecto de la imagen original) de alto han proporcionado los mejores resultados.
- *Reducción de la dimensionalidad*. Se ha aplicado una transformación PCA a los vectores de características locales y una reducción de dimensionalidad. La técnica PCA solamente se ha aplicado a las componentes del vector de caracteristicas resultantes de concatenar los valores de gris de los píxeles de las ventanas, es decir, no se han tenido en cuenta las dos componentes de posición de la ventana. Se han probado reducciones entre 10 y 25 dimensiones, obteniendo los mejores resultados en cuanto a coste computacional y tasa de error con las primeras 15 componentes PCA. Esto significa que los vectores de características con los que se trabajará finalmente tendrán 17 componentes: las 15 de la reducción PCA más las 2 coordenadas de posición de la ventana.
- *Peso de las características globales*. Después de la reducción de la dimensionalidad, se han agregado a cada vector de características las coordenadas de posición del punto central de cada ventana. Los valores de estas nuevas características se han normalizado a la desviación estándar del primer componente PCA y se han multiplicado por un factor de peso para sintonizar su efecto con respecto al resto de los componentes. Se han probado combinaciones de factores de peso (α_x, α_y) entre 0 y 8.
- *Paso de submuestreo y orla de vecindad*. Se han efectuado varios experimentos con estos dos parámetros. Finalmente, se ha llegado a la conclusión de que estos parámetros están fuertemente ligados al tamaño de la ventana de características locales. Los mejores resultados se obtienen aplicando pasos de submuestreo y orla de vecindad con valores de la mitad del tamaño de la ventana

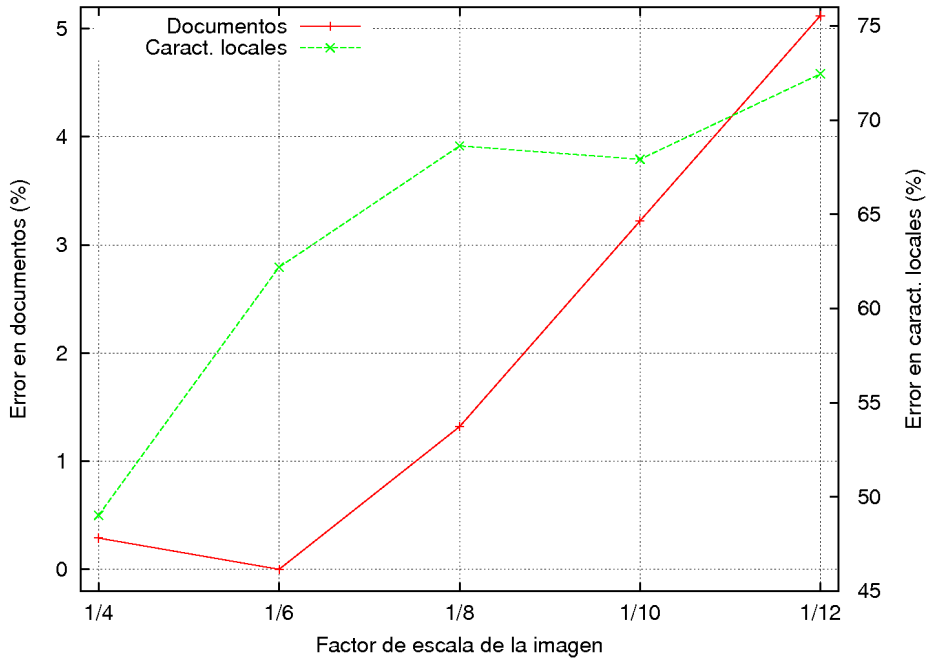


Figura 5.1: Tasas de error a nivel de documento (eje izquierdo) y nivel de característica local (eje derecho) para diferentes escalas de reducción. Se muestran los datos logrados con la mejor combinación del resto de parámetros.

de características. P. ej.: para una ventana de tamaño 14×8, se utilizarán pasos de submuestreo $(s_x, s_y) = (7, 4)$ y un tamaño de orla de vecindad de 7×4

- *Numero de características locales (LF)*. Se han seleccionado distintas cantidades de subimágenes que presentan los mejores índices de contraste (máxima varianza) para los conjuntos de entrenamiento y test. Se han probado valores entre 100 y 1000. Para reducir el número de cálculos en la búsqueda de candidatos, se ha aplicado submuestreo a las imágenes, tanto en la fase de entrenamiento como en la de test. En la fase de entrenamiento, se han incluido en los conjuntos de cada clase las subimágenes extraidas de la orla de vecindad de cada píxel seleccionado.

Se ha implementado un clasificador rápido basado en la técnica de búsqueda aproximada mediante *kd*-tree. Se ha obtenido el vecino más próximo de cada uno de los vectores 17-dimensionales (15 + 2) extraidos de las imágenes de test, utilizando un valor de $\epsilon = 2$, y se ha empleado la *"regla de la suma"* descrita en el apartado 2.3 para clasificar estas imágenes.

Se ha alcanzado un resultado óptimo de 0 % de tasa de error en la clasificación de documentos con la siguiente combinación de parámetros:

- Dimensión del vector PCA: 15 componentes
- Tamaño de la ventana de características (antes de escalado): 80×30
- Submuestreo (antes de escalado): 40×15
- Orla de vecindad (antes de escalado): 40×15
- Factor de escala: $1/6$

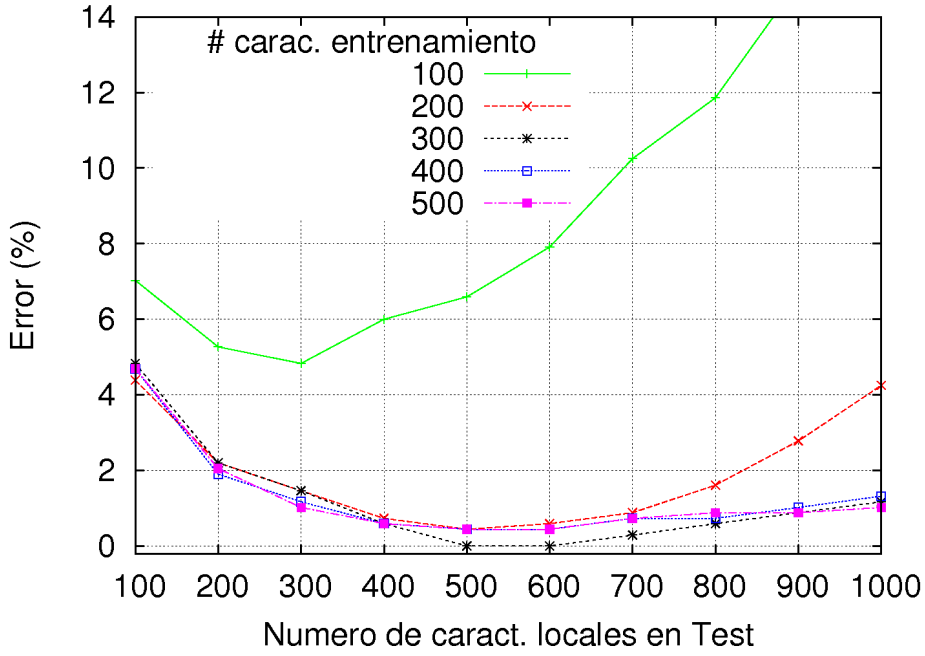


Figura 5.2: Tasa de error a nivel de documento en función del número de características locales utilizadas en test y entrenamiento.

- 300 vectores de entrenamiento por clase
- 500 vectores de test
- Peso de las coordenadas $(\alpha_x, \alpha_y) = (6, 2)$

El tiempo promedio para la identificación (fase de test) ha sido de 4,5 documentos/s en un ordenador con procesador AMD de 64 bits y 4 CPU de 3 GHz.

La figura 5.1 muestra el error de clasificación obtenido para los distintos valores de escala del documento, en cada valor obtenido se ha utilizado la mejor combinación del resto de parámetros.

El número de las características locales y el peso de las características globales (coordenadas de posición de los vectores) son parámetros fuertemente relacionados con la aproximación utilizada. Por esto, en la gráfica 5.1 se presenta un análisis de los resultados sobre la variación de ambos parámetros, al mismo tiempo que se fijan los restantes. Se observa un resultado óptimo con una tasa de error en la clasificación de documentos del 0 % para una reducción de escala de 1/6. La curva de color rojo representa la tasa de error de clasificación de las subimágenes seleccionadas. En cada punto representado se han fijado el resto de parámetros a los valores que producen un resultado óptimo.

En la figura 5.2 se muestra el error obtenido en función del número de vectores seleccionados en las fases de entrenamiento y test. Generalmente, se han encontrado las mejores combinaciones utilizando unos cuantos vectores más para la fase de test que los seleccionados en la fase de entrenamiento. No se aprecian grandes diferencias entre las distintas combinaciones. Aunque se produce el resultado óptimo de 0 % de error en la combinación de 300 vectores seleccionados en fase de entrenamiento y 500 en fase de test.

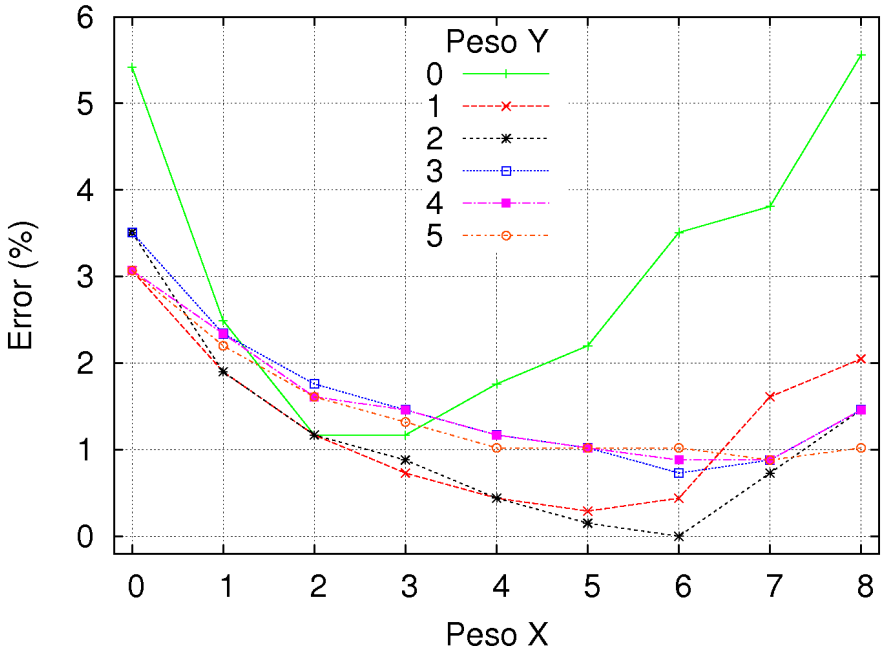


Figura 5.3: Tasa de error a nivel de documento en función de los pesos de las coordenadas.

La figura 5.3 muestra la gran capacidad discriminante que aporta la información de las coordenadas de posición de las subimágenes. La combinación de pesos $(\alpha_x, \alpha_y) = (6, 2)$ produce el resultado de 0 % de tasa de error en la clasificación de documentos. Esto significa una mejora del 5 % (33 documentos) respecto de los resultados obtenidos sin tener en cuenta la posición como una característica global, valor que se puede observar en el punto $(\alpha_x, \alpha_y) = (0, 0)$ de la gráfica. Cabe destacar la distinta influencia en los resultados de la coordenada x, respecto de la coordenada y, que es mucho menos discriminante que la primera. Esto es debido probablemente a que los documentos presentan traslaciones más acusadas en la coordenada y debido a los defectos mecánicos en el proceso de escaneado.

5.4. Opción de rechazo

Un buen clasificador debe ser robusto ante la entrada de imágenes que no pertenecen a ninguna de las clases conocidas, debe ser capaz de detectar estos documentos y rechazarlos. Para ello se ha establecido una medida de calidad de la clasificación de un documento o "índice de fiabilidad", de esta forma, se establecerá un umbral para este índice de fiabilidad y se rechazarán aquellos documentos que no lo superen.

Para un conjunto de test dado, la distribución de los índices de fiabilidad de los documentos bien clasificados no deberia solaparse con la distribución de los índices de fiabilidad de documentos desconocidos o mal clasificados. Obviamente, cuanto mayor sea la separación entre ambas distribuciones se esperará una mejor capacidad de generalización.

Se han seleccionado 200 documentos aleatorios correspondientes al corpus descrito en el apartado 5.1 y se han añadido al conjunto de test como una clase especial de "documentos desconocidos". Se ha efectuado una clasificación con los parámetros óptimos mostrados en el apartado 5.3 y se han obtenido

los votos recibidos en cada clases para todos los documentos de test.

Para un documento cualquiera de test, se ha definido una función de fiabilidad basada en las dos clases más votadas de la siguiente forma:

$$F = \alpha f + (1 - \alpha)g$$

siendo:

- $f = f^1$ la probilidad a posteriori de la clase más votada, con $f^1 = \frac{n_1}{N}$, donde n_1 es el número de votos obtenidos por la clase más votada y N el número de características extraidas en un documento.
- $g = f^1 - f^2$ es la diferencia de probabilidades a posteriori entre las dos clases más votadas.
- α es un valor entre 0 y 1 con el que se ha experimentado.

Los experimentos han demostrado que cualquier valor de α ofrecía los mismos resultados, por lo que la función de fiabilidad se ha simplificado para el valor de $\alpha = 0$, es decir, la función de fiabilidad empleada finalmente ha sido:

$$F = f^1 = \frac{n_1}{N}$$

Aplicando este criterio como índice de fiabilidad a los resultados de clasificación del conjunto original de test junto con las imágenes de documentos desconocidos se ha obtenido un valor del 99,85 % de exhaustividad al 100 % de precisión (*recall at 100 % precision*) 2.5.1.

Con estos resultados, se puede concluir afirmando que este sistema será capaz de aceptar con un elevado porcentaje de acierto los documentos conocidos y bien clasificados, y de rechazar los documentos desconocidos y los que han sido clasificados de forma incorrecta.

Capítulo 6

Conclusiones

En el presente trabajo se ha presentado la aplicación de una técnica para la identificación de imágenes de documentos con información cumplimentada utilizando una combinación del uso de características locales junto con un sistema de clasificación basado en un esquema de votación directa sobre un clasificador de los *k*-vecinos más próximos.

Se ha recopilado y preparado una base de datos extensa de imágenes de documentos con la que se han realizado los experimentos, y que puede servir para futuros trabajos relacionados con documentos con información cumplimentada.

Los resultados obtenidos son representativos de los que se deberían obtener en un proceso real de similares características. La base de datos de documentos se ha recopilado a partir de facturas, formularios, recibos, y otros tipos de documentos, escaneados mecánicamente en un proceso comparable al que seguiría el *workflow* habitual de una empresa con necesidad de digitalización de documentos.

Se ha realizado una experimentación exhaustiva de los parámetros más relevantes, identificando sus valores óptimos, con los cuales se ha alcanzado una tasa de error del 0 % en la clasificación de documentos, y un resultado del 99,85 % de *recall at 100 % precision* al incluir la opción de rechazo de documentos desconocidos.

Los tiempos de proceso requeridos (alrededor de 4,5 documentos por segundo) hacen pensar que este sistema sería fácilmente integrable en un aplicación real de identificación de documentos, incluso para empresas con elevado *workflow*.

Bibliografía

[Andreu-Cerezo 10] Luis Andreu-Cerezo. Detección y modelización de casillas de campos de formularios. Master's thesis, Universitat Politècnica de València, 2010.

[Arlandis 03] Joaquim Arlandis. *La transformació contínua de la distància. Estudi i aplicació a un sistema OCR*. PhD thesis, Universitat Politècnica de València, 2003.

[Arlandis 09] Joaquim Arlandis, Juan-Carlos Perez-Cortes & Emilio Ungria. *Identification of very similar filled-in forms with a reject option*. In ICDAR, pages 246–250, 2009.

[Arlandis 11] Joaquim Arlandis, Vicent Castello-Fos & Juan-Carlos Perez-Cortes. *Filled-in document identification using local features and a direct voting scheme*. In IbPRIA, 2011.

[Dimmick 92] D. L. Dimmick & M. D. Garris. *Structured Forms Database 2, NIST Special Database 6*. Rapport technique, National Institute of Standards and Technology, 1992.

[Doermann 98] David Doermann. *The Indexing and Retrieval of Document Images: A Survey*. Computer Vision and Image Understanding, vol. 70, no. 3, pages 287 – 298, 1998.

[Díaz 03] Raquel Gómez Díaz. *La evaluación en recuperación de la información*. Hipertext.net, vol. 1, 2003.

[Fan 01] Kuo-Chin Fan, Mei-Lin Chang & Yuan-Kai Wang. *Form Document Identification Using Line Structure Based Features*. In ICDAR, pages 704–708, 2001.

[Haralick 73] Robert M. Haralick, K. Shanmugan & Its'Hak Dinstein. *Textural Features for Image Classification*. IEEE Transactions on Systems, Man, and Cybernetics, vol. 3, pages 610–621, 1973.

[Heroux 98] Pierre Heroux, Sebastien Diana, Arnaud Ribert & Eric Trupin. *Classification Method Study for Automatic Form Class Identification*. In Proc. 14th Int. Conf. on Pattern Recognition, ICPR'98, pages 926–928, 1998.

[Kittler 98] Josef Kittler, Mohamad Hatef, Robert P. W. Duin & Jiri Matas. *On Combining Classifiers*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 3, pages 226–239, 1998.

[Mandal 05] S. Mandal, S. P. Chowdhury, A. K. Das & Bhabatosh Chanda. *A Hierarchical Method for Automated Identification and Segmentation of Forms*. Document Analysis and Recognition, International Conference on, vol. 0, pages 705–709, 2005.

[Mohr 97] Roger Mohr, S. Picard & Cordelia Schmid. *Bayesian Decision Versus Voting for Image Retrieval*. In CAIP, pages 376–383, 1997.

[Nagasaki 06] Takeshi Nagasaki, Katsumi Marukawa, Tatsuhiko Kagehiro & Hiroshi Sako. *A Coupon Classification Method Based on Adaptive Image Vector Matching*. In 18th International Conference on Pattern Recognition, pages 280–283, 2006.

[Ogata 03] Hisao Ogata, Shigeru Watanabe, Atsuhiro Imaizumi, Tsukasa Yasue, Naohiro Furukawa, Hiroshi Sako & Hiromichi Fujisawa. *Form-type identification for banking applications and its implementation issues*. In DRR, pages 208–218, 2003.

[Ohtera 04] Ryo Ohtera & Takahiko Horiuchi. *Faxed Form Identification using Histogram of the Hough-Space*. Pattern Recognition, International Conference on, vol. 2, pages 566–569, 2004.

[Paredes 01] Roberto Paredes, Juan Carlos Pérez-Cortes, Alfons Juan & Enrique Vidal. *Local Representations and a direct Voting Scheme for Face Recognition*. In PRIS, pages 71–79, 2001.

[Parker 10] Charles Parker. *Anchor point selection by KL-divergence*. In Image Processing Workshop (WNYIPW), pages 42 – 45. IEEE Computer Society, 2010.

[Sako 03] Hiroshi Sako, Minenobu Seki, Naohiro Furukawa, Hisashi Ikeda & Atsuhiro Imaizumi. *Form Reading based on Form-type Identification and Form-data Recognition*. Document Analysis and Recognition, International Conference on, vol. 2, page 926, 2003.

[Sarkar 06] Prateek Sarkar. *Image classification: Classifying distributions of visual features*. Pattern Recognition, International Conference on, vol. 2, pages 472–475, 2006.

[Sarkar 10] Prateek Sarkar. *Learning Image Anchor Templates for Document Classification and Data Extraction*. Pattern Recognition, International Conference on, vol. 0, pages 3428–3431, 2010.

[Ting 96] A. Ting & M. Leung. *Business Form Classification Using Strings*. In ICPR96, pages II: 690–694, 1996.