Algunas aportaciones a la segmentación de imagen digital basadas en el Mapa Autoorganizativo

Imanol Echave Calvo

The Date

Resumen

La presente tesis se enmarca en la frontera entre dos disciplinas y métodologías básicas para atacar los problemas de tratamiento de señal y, específicamente, de tratamiento de imagen. La primera está basada en métodos adaptativos de minimización del error cuadrático. La segunda plantea del modelado bayesiano de la imagen y trata de obtener la estimación de máxima probabilidad *a posteriori* de la imagen procesada. En el primer caso las técnicas numéricas aplicadas son técnicas de minimización mediante descenso del gradiente, en el segundo se aplican técnicas de relajación estadística (*simulated annealing*) o simplificaciones que dan respuestas locales al problema de minimización global planteado. En esta contexto, proponemos la utilización de los métodos de clustering neuronales no supervisados (el algoritmo SOM en particular, como técnicas de preproceso de la imagen en un marco bayesiano. Justificamos este proceso y lo denominamos Filtrado Bayesiano basado en la Cuantización Vectorial (VQ-BF). Finalmente proponemos la aplicación en dos campos dispares: la clasificación no supervisada de imágenes de resonancia magnética (MRI) y el cálculo del flujo óptico. Para que esta aplicación sea factible se precisa de un algoritmo de clustering rápido para evitar los tiempos de respuesta fuera de los límites que imponen las aplicaciones en tiempo real. Para este propósito estudiamos las diversas arquitecturas competitivas neuronales que encontramos en la literatura y las probamos empíricamente en un contexto de un problema de cuantización del color de la imagen. Los resultados empíricos apoyan al Mapa Auto-Organizativo (SOM) como la mejor alternativa. Finalmente, consideramos el problema de la determinación del número de clases en la clasificación no supervisada bajo la perspectiva de los denominados "filtros de Occam". En esta metodología de filtrado se utilizan algoritmos de compresión para la eliminación del ruido impulsivo aditivo. La estimación de la magnitud del ruido se realiza estudiando la curva de la razón de compresión versus la distorsión después de la compresión. En nuestro caso, el algoritmo de compresión es la cuantización vectorial realizada mediante el libro de código obtenido por la arquitectura competitiva.

Índice General

Índice de Figuras

Índice de Tablas

1. Introducción

La literatura sobre proceso de imagen y sobre redes neuronales es extensísima, por lo que toda nueva aportación es usualmente muy matizada y local. Si se nos permite la metáfora culinaria, los ingredientes de tesis son de dos clases: los problemas a resolver y las técnicas numéricas utilizadas. En esta tesis doctoral se desarrollan aplicaciones de procesado de la imagen digital basada en el algoritmo del Mapa Auto-organizativo (*Self-Organizing Map*) (SOM). En particular se considera la posibilidad de utilizar el SOM como herramienta para la estimación de patrones de bloques de las imágenes para su filtrado posterior. Al proceso lo hemos denominado Filtrado Bayesiano basado en la Cuantización Vectorial (VQ-BF) y lo hemos aplicado en dos campos radicalmente distintos: el cálculo del flujo óptico en aplicaciones de visión por computador y el procesado de imágenes de resonancia magnética. En este capítulo de introducción revisaremos brevemente ambos aspectos de la tesis antes de proporcionar una visión más sumaria y esquemática dada por la especificación de objetivos, estructura de la tesis y publicaciones realizadas durante el proceso de desarrollo de los trabajos.

1.1.Las aplicaciones

Las dos aplicaciones en las que hemos centrado nuestra atención son al nivel de la segmentación de la imagen. Como es bien conocido de la literatura general sobre proceso digital de imagen y visión por computador [28], [47], [116], la estructura de un sistema de visión por computador se descompone en las siguientes etapas:

- Captura: se obtiene la imagen mediante algún dispositivo sensor, en nuestro caso cámaras ópticas convencionales.

- Preproceso: se transforma la imagen mediante algoritmos numéricos para eliminar los diversos efectos de la incertidumbre, que en proceso de señal se agrupan genéricamente bajo el nombre de ruido. La incertidumbre espacial se corrige mediante algoritmos de registro, el ruido térmico que produce

ruido impulsivo aditivo se corrige mediante filtrado frecuencial o espacial, las variaciones de iluminación se corrigen mediante la estimación del campo de iluminación, y un largo etctera.

- Segmentación: la segmentación produce la partición de la imagen en regiones que se espera tengan un significado específico. Esta partición puede realizarse atendiendo a los contornos de las cosas mediante el analisis de bordes o atendiendo a las agrupaciones de píxeles similares en regiones espaciales según criterios de crecimiento de regiones o de texturas.

- Reconocimiento e interpretación: en muchos casos el objetivo final de los procesos de imagen es el reconocimiento de alguún objeto presente en la imagen o el eitquetado de los píxeles de la imagen para que sena interpretados por el operador humano. A veces parte de este proceso de reconocimiento consiste en la construcción de modelos geométricos de la realidad a partir de las imágenes captadas.

1.1.1.La imagen de resonancia magnética (MRI)

Las imágenes de resonancia magnética (MRI) permiten visualizar con gran contraste los tejidos blandos y ha revolucionado la capacidad de diagnosticar las patologías que los afectan. La visualización de resonancia magnética está basada en el fenómeno conocido como resonancia magnética nuclear (RMN) y las MRI son mediciones agregadas de la composición de los tejidos al nivel molecular. El hecho de que la estructura molecular permanece constante dentro del tejido y varía entre distintos tejidos determina la efectividad de la MRI. La MRI tiene una alta resolución espacial y proporciona mucha información sobre la estructura anatómica, permitiendo estudios patológicos o clínicos cuantitativos, la derivación de atlas anatómicos digitalizados y también la guía antes y durante la intervención terapéutica. Las aplicaciones iniciales de la MRI requirieron principalmente que la imagen fuera lo suficientemente clara y libre de artefactos como para soportar las tareas de diagnóstico cualitativo de patologías.Muchas de las tareas que se requieren de los radiólogos son repetitivas, lo que justifica la introducción de métodos automatizados de proceso de la imagen. Por ello existe una necesidad creciente de resultados cuantitativos basados en el análisis automatizado de la imagen. El proceso automatizado puede ir desde la eliminación de ruido impulsivo, mediante filtros frecuenciales y otros como los filtros anisotrópicos, hasta las operaciones de registro y segmentación. El registro de las imágenes consiste

en el alineamiento de diferentes imágenes para obtener una visión más completa. Se aplica en la mezcla de imágenes de múltiples modalidades y la comparación de los pacientes con atlas anatómicos. La segmentación de la imagen consiste en su descomposición en regiones. Los criterios de identificación de regiones varían de aplicación en aplicación. La segmentación de la imagen es crítica en aplicaciones como la diagnosis de la esquizofrenia, la detección de tumores y la cirugía basada en realidad aumentada. Los métodos de segmentación deben ser fiables y reproducibles. Por ejemplo, en estudios de enfermedades degenerativas cerebrales como la esquizofrenia, el mal de Alzheimer [126], esclerosis multiple, es necesario medir de forma precisa la cantidad de materia gris, materia blanca, lesiones en la materia blanca, líquido cerebro-espinal y sus distribuciones espaciales y cambios temporales. Las imágenes de resonancia magnética presentan dificultades específicas para su proceso y segmentación. Son imágenes de muy bajo contraste entre tejidos con fuerte ruido que puede ser debido a diversas circunstancias de la captura de la imagen. Por ejemplo, el ruido aditivo, que sigue una distribución Rician, es producto tanto de las limitaciones en la precisión numérica, como del ruido térmico en la captura de la señal por parte de la antena emisora-receptora. La descripción completa y detallada de los procesos de captura de la MRI y de los diversos artefactos que pueden afectarla se pueden encontrar en [25], [31], [59], [88], [112], [128]. En la literatura se han aplicado métodos automáticos y semi-automáticos para la segmentación y la medición volumétrica. En nuestro caso hemos propuesto un método híbrido, con una fase de entrenamiento supervisada y otra, previa, no supervisada. Sobre las imágenes de resonancia magnética (MRI) hemos realizado experimentos de filtrado, segmentación no supervisada y de clasificación supervisada. Todos los experimentos se han realizado sobre imágenes cedidas por la antigua Unidad de Resonancia Magnética del Instituto Pluridisciplinar de la Universidad Complutense de Madrid, actualmente Instituto de Estudios Biofuncionales. La aplicación fundamental ha sido la de clasificación supervisada para la detectar zonas infecciosas en un experimento de seguimiento de la evolución de un modelo de infección por la bateria *Aspergillus Fumigatus*. También se han realizado experimentos para la segmentación del *Corpus Callosum* en imágenes de resonancia magnética del cerebro.

1.1.2.El cálculo del flujo óptico

El movimiento de los objetos en el mundo o el movimiento de la cámara producen movimientos en la imagen. El flujo óptico es el movimiento aparente en el plano

imagen producido por el movimiento real en el mundo. Este movimiento aparente puede ser utilizado de diversas maneras para diversas aplicaciones que van desde el seguimiento de objetos, las estimación de modelos tridimensionales, hasta la autolocalización de agentes autónomos. Las variaciones en la imagen pueden deberse también a otros fenómenos como los cambios de iluminación o el ruido de captura de la cámara. Los algoritmos de seguimiento de objetos, por ejemplo, deben ser capaces de distinguir estos cambios en la imagen de los relacionados con movimientos de objetos en el mundo. El cálculo del flujo óptico es, po tanto, una de las tareas clásicas en visión por computador [63], [58].

El algoritmo más clásico de cálculo del flujo de movimiento es el propuesto por Horn [63]. Se basa en la constancia de la intensidad de la imagen, esto es en la idea de que la intensidad es constante si corregimos los movimientos que se han producido en la escena. El algoritmo es básicamente a nivel de pixel y tiene el inconveniente de que las superficies suaves, casi constantes, no producen detección del movimiento excepto en las fronteras de los objetos. Lo que se conoce como el efecto de la apertura. Interesan otras soluciones más robustas y de mayor eficiencia en tiempo de cálculo. Otras aproximaciones, que se revisan en el capítulo 5 están basadas en el modelado bayesiano de la imagen y la detección mediante correlación. Cómo se verá, el algoritmo propuesto está relacionado más bien con estos últimos algorithmos.

Una de las características de interés de los algoritmos de cálculo del flujo óptico es si producen o no estimaciónes densas del flujo. La estimación es densa si tenemos un vector de movimiento para cada pixel de la imagen. Las estimaciónes densas pueden ser más deseadas para algunas aplicaciones como el seguimiento de objetos, donde se pueden segmentar los objetos en la imagen en función de los vectores de movimiento de los pixels. Debido a la importante carga de cálculo, se han realizado Númerosas proposiciones de algoritmos basadas en puntos especiales de la imagen y contornos de los objetos dando lugar a estimaciónes dispersas (sparse) del flujo. Sin embargo, nosotros estamos interesados en estimaciónes densas en las que para cada pixel de la imagen tenemos el vector de su desplazamiento aparente en la imagen. El algoritmo propuesto, en breve resumen, consiste en la segmentación en regiones de las imágenes y el cálculo de las correspondencias de las regiones estimadas en las imágenes consecutivas.

1.1.3.La cuantización del color

La cuantización del color es el proceso de obtener una paleta de colores o un conjunto de colores representantes de una imagen. En su origen [61] el problema se planteaba como un problema de visualización. Dado que los monitores de los ordenadores tenían una memoria limitada se representaban las imágenes como matrices cuyos elementos eran índices a una tabla de colores. Estas limitaciones desaparecieron a principios de la década de los años 90, sin embargo el interés por los algoritmos de cuantización del color no desapareció sino que se transformó en una herramienta para el análisis y segmentación de la imagen. Por ejemplo, muchos algoritmos de localización de objetos (i.e. caras) se basan en la obtención de regiones de color definidas a partir de procesos de cuantización de color. Estas regiones pueden ser fácilmente identificadas y correladas entre imágenes sucesivas de una secuencia. En esta tesis, la cuantización del color es un problema benchmarck para la evaluación dela calidad relativa de los distintos algoritmos competitivos.

1.2.Los útiles matemáticos

El algoritmo VQ-BF propuesto, puede exponerse de forma muy resumida como el procesado de cada pixel en función de su vecindario comparandolo contra una paleta de bloques de las imágenes que ha sido estimada de la imagen o de una colección de imágenes similares mediante algorimos de agrupamiento, en este caso el SOM. En el caso de la segmentación no supervisada el algoritmo VQ-BF se encarga de estimar las posibles clases de los pixels a partir de la clasificación no supervisada (clustering) de los bloques en la imagen, entendidos como vecindarios de los píxeles. En la clasificación supervisada, el VQ-BF realiza el papel de un preproceso que facilita la clasificación supervisada realizada por métodos típicos (i.e. redes neuronales artificiales). El VQ-BF puede aplicarse también como método de filtrado si el valor de intensidad del pixel se sustituye por el valor correspondiente al pixel central del bloque más similar al vecindario del pixel. En resumen, los recursos matemáticos involucrados son algoritmos de agrupamiento (clustering) realizados mediante algoritmos de redes neuronales competitivas y el modelado bayesiano de la imagen.

1.2.1.Redes neuronales competitivas

En el área general de las redes neuronales artificiales, las redes neuronales competitivas son algoritmos dedicados a la clasificación no supervisada, esto es, a resolver los problemas de agrupamiento de datos basados en representantes. Los representantes son los pesos de las conexiones en terminología neuronal. Las redes neuronales artificiales se conceptuan y justifican desde diversos puntos de vista. Son, en muchos casos, algoritmos adaptativos online de minimización del error cuadrático. En otras ocasiones [1] se presentan como implementaciones altamente paralelas alternativas a las implementaciones secuenciales convencionales, sin embargo esta perspectiva se está abandonando progresivamente. En el caso de las redes neuronales competitivas, se trata de algoritmos de minimización de una determinada función de distorsión, que varía de un esquema de red a otro. En el capítulo 2 presentamos distintas alternativas de redes neuronales competitivas en este marco conceptual.

Las propiedades de convergencia de estas redes son muy pobres, esto es, necesitan largos periodos de entrenamiento para obtener los resultados deseados, al menos en la teoría. Nuestras necesidades exigen respuestas en tiempo breve, por lo que es necesario considerar alternativas rápidas de entrenamiento. Exploramos la posibilidad y validez de los entrenamientos en una única pasada sobre la muestra de los datos.

También se puede considerar las redes neuronales desde el punto de vista de la teoría de la decisión y clasificación bayesiana. Bajo este punto de vista son algoritmos de clasificación no supervisada. En el caso más sencillo, el algoritmo competitivo simple, se trata de un algoritmo de clasificación de máxima probabilidad *a posteriori* asumiendo que la distribución de probabilidad de los datos es una mezcla de distribuciones gausianas, y el algoritmo de entrenamiento es una versión online del algoritmo de estimación de máxima verosimilitud, coincidiendo también con el algoritmo de maximización de la expectación (*Expectation Maximization* EM). En los casos más sofisticados, no está claro cual es el modelo de la distribución de probabilidad de los datos y los algoritmos de entrenamiento son variaciones sobre el algoritmo competitivo simple.

1.2.2.Modelado bayesiano de la imagen

Se presenta en los trabajos de [41] que introducían el *simulated annealing* como algoritmo estocástico de optimización global. Formula las operaciones de transformación de las imágenes, incluidas las de clasificación a nivel de pixel, co-

mo estimaciónes de máxima probabilidad a posteriori en un marco de modelado bayesiano. Usualmente, el modelo de la probabilidad condicionada corresponde al modelo de ruido aditivo y el modelo de la probabilidad *a priori* corresponde con condiciones de regularización de la solución, por ejemplo condiciones de suavidad de la imagen. El proceso habitual consiste en formular el problema especificando los modelos de probabilidad *a priori* y condicional, estos modelos son habitualmente exponenciales, bien gausianos bien modelos markovianos, los denominados Markov Random Field (MRF) que proporcionan modelos espaciales de la intensidad de la imagen. La expresión de la probabilidad *a posteriori* toma la forma de una distribución exponencial, la distribución de Gibbss o Boltzmann en el caso clásico, y su exponente se asocia a una función de energía. La estimación de máxima probabilidad *a posteriori* se realiza aplicando algún algoritmo de relajación o búsqueda aleatoria, el más clásico el *simulated annealing*, a la minimización de la función de energía.

En nuestro caso, el algoritmo que definimos se asocia a una decisión de máxima probabilidad y lo que tratamos de deducir es el modelo *a priori*, asumiendo un modelo gausiano del ruido. Esto es, se realiza el camino en el sentido opuesto al convencional. Lo que nos dice el modelo *a priori* es el grado de robustez del que está dotado el algoritmo.

1.3.Objetivos iniciales de la tesis y grado de consecución

Los obejtivos generales al momento de plantear los trabajos de la tesis eran:

- La obtención de un algoritmo de segmentación de imágenes robusto y eficiente en términos de tiempo de cálculo. Este es un objetivo sumamente genérico que en esencia plantea el marco global de la tesis, sitúa el área de trabajo.

- La obtención de métodos de segmentación de imágenes de resonancia magnética (MRI). Las imágenes de MRI se caracterizan por tener un bajo contraste y ser muy ruidosas, si bien el estado del arte de los sistemas de captura ha mejorado considerablemente. Estas características de la imagen exigen que los algoritmos de segmentación sean al mismo tiempo extremadamente robustos al ruido y muy sensibles para poder discriminar tejidos con poco contraste. Hemos logrado obtener algoritmos novedosos y que presentas algunas propiedades de robustez muy interesantes. Son dos tipos de algoritmos esencialmente distintos: los no supervisados que procesan la imagen

utilizando sólo los resultados de los algoritmos de clustering, y los híbridos que utilizan aplican algoritmos de entrenamiento supervisados sobre los resultados del procesado no supervisado.

- La obtención de métodos robustos de cálculo del flujo óptico. El flujo óptico es la herramienta básica para la observación del movimiento en las imágenes y secuencias de imágenes. Es una herramienta potencialmente muy útil en aplicaciones diversas, como las robóticas y las de interacción persona computador. Los métodos convencionales son muy sensibles al ruido y a los cambios de iluminación. Nuestro objetivo era la definición de un algoritmo basado en la descomposición en regiones, bajo la hipótesis de que dichas regiones se mantendrán estables a lo largo de la secuencia de imágenes. La aplicación de nuestros algoritmos de proceso basados en el SOM produce efectivamente regiones bastante estables. Los cambios bruscos de iluminación producen efectos no recuperables y los gradientes en las grandes superficies pueden también ser fuente de artefactos extraños.

Nuestro grupo de trabajo tenía una cierta tradición en la aplicación de redes competitivas, en especial el SOM, al momento de comenzar los trabajos de la tesis doctoral, de ahí que nuestro interés inicial se concentrase en este tipo de algoritmos como herramientas computacionales. En relación con ellos nos planteamos los siguientes objetivos

- Determinar la posibilidad de aplicación rápida de los algoritmos. Para la aplicación en situaciones de tiempo real, es preciso que la carga computacional del algoritmo no sea excesiva y permanezca acotada. Las redes neuronales y otros algoritmos similares se caracterizan por un tiempo de convergencia muy lento de acuerdo a las condiciones impuestas por el algoritmo de descenso de gradiente estocástico. Nos proponíamos evaluar la pérdida de calidad que produce una realización en un sólo paso sobre la muestra. Hemos realizado un estudio exhaustivo en un dominio concreto, el de la cuantización del color. Estos resultados no tienen una validez universal pero creemos que son un buena orientación para estudios y trabajos posteriores. Los trabajos desde un punto de vista analítico están en curso en el contexto de otras tesis doctorales de nuestro grupo de trabajo y se orientan al estudio de las propiedades analíticas de los algoritmos competitivos como algoritmos de continuación de puntos de equilibrio y como algoritmos de convexidad gradual.

- Definir la forma en que un algoritmo neuronal competitivo puede ser utilizado en el proceso de la imagen para propósitos de filtrado y/o clasificación. En el caso de las imágenes en color, la segmentación basada en el color se formula inmediatamente como un proceso de agrupamiento estadístico (*clustering*), esto no es tan inmediato en el caso de que se pretenda procesar grupos o bloques de píxeles. En este sentido propusimos y probamos dos esquemas de tratamiento de las imágenes en los que los bloques de píxeles juegan un papel de elementos de textura que se aprenden a través del algoritmo de clustering, cualquiera que sea este.

- Intentar justificar formalmente los algoritmos de proceso definidos. El marco más general para la interpretación de los procesos realizados sobre las imágenes nos parece que es el marco bayesiano. La pregunta a responder es ¿cual es la interpretación bayesiana de un proceso realizado siguiendo un determinado algoritmo? En este sentido, el razonamiento bayesiano no sirve para la definición de un nuevo algoritmo sino para la justificación e interpretación de uno dado. La definición formal del VQ-BF y la deducción de sus propiedades se enmarcan en este proceso.

- Estudiar el problema de la determinación del número de clases. Este es uno de los problemas clásicos en los problemas de clasificación no supervisada. En el contexto de la presente tesis, este problema se concreta en la determinación del número de clases de textura que bastarán para modelar las características de la imagen. Para atacar este problema nos inspiramos en los denominados algoritmos de Occam y conseguimos algunos resultados preliminares en el tratamiento de la imagen de MRI.

1.4.Estructura de la memoria de la tesis

La memoria se estructura en dos partes bien diferenciadas, la primera corresponde a la presentación de las arquitecturas competitivas, incluido el SOM y su ejecución en un único paso sobre la muestra como una alternativa viable computacionalmente y con buenos y sorprendentes resultados. Parte de este trabajo se ha realizado en colaboración con Ana Isabel Gonzalez de Acuña y nos hemos apoyado en algunos de sus resultados recientes.En la primera parte de la memoria se incluye la definición general del algorimo VQ-BF y su análisis dentro del marco del procesado bayesiano de las imágenes. La segunda parte de la memoria corre-

sponde con la descripción de las dos aplicaciones probadas y la presentación de los resultados.

El capítulo 2 se introduce las redes neuronales competitivas y algoritmos de agrupamiento adaptativos similares, en el marco clásico de la teoría bayesiana de la decisión, mostrando que se trata de esquemas de descenso de gradiente.

El capítulo 3 presenta la justificación empírica de la utilización de un algoritmo de entrenamiento con un solo paso sobre la muestra, en el marco de un problema de cuantificación del color. Se concluye con la recomendación del SOM como el mejor en términos de tiempo de respuesta y calidad.

El capítulo 4 presenta el filtrado bayesiano de la imagen utilizando representantes de bloques de píxeles, que se corresponden con los libros de códigos obtenidos mediante algoritmos de agrupamiento adaptativo. Denominamos VQ-BF a este algoritmo.

El capítulo 5 presenta la aplicación del filtrado bayesiano basado en la cuantización vectorial (VQ-BF) para el cálculo del flujo óptico en secuencias de imágenes.

El capítulo6 presenta la aplicación del filtrado bayesiano basado en la cuantización vectorial (VQ-BF) a la segmentación de imágenes de resonancia magnétic nuclear, en dos casos concretos: la medida de un proceso inflamatorio y la medición de estructuras del cerebro humano.

El capítulo 7 presenta una aplicación de los filtros de Occam para la determinación del número óptimo de vectores código en el caso de la segmentación no supervisada de imágenes de resonancia magnética, más precisamente de la imagen de un embrión humano.

El apendice A revisa el método del gradiente estocástico en el que se basan las redes neuronales competitivas.

El apéndice B recoge la prueba de la convergencia topológica del SOM en el caso escalar.

EL apéndice C recoge la definición de la metodología de los filtros de Occam.

El apéndice D recoge la derivación formal de la regla adaptativa FLVQ.

1.5.Publicaciones realizadas durante el desarrollo de la tesis

Se han realizado o participado en su realización, durante el desarrollo de estas tesis las siguientes publicaciones en obras colectivas, usualemente procedentes de congresos:

- Fast face localization for mobile robots: signature analysis and color processing, M. Graña, A. I. Gonzalez, B. Raducanu, I. Echave; en Intelligent robots

and computer vision XVII: Algorithms, Techniques and Active Vision, D. P. Casasent (ed) SPIE Conference Proceedings. Vol. 3522 pp.387-398

- Bayesian VQ image filtering design with fast adaptation competitive neural networks;A. I. Gonzalez, M. Graña, I. Echave, J. Ruiz-Cabello; IWANN99, Alicante, Junio 1999, en Engineering Applications of Bio-inspired Artificial Neural Networks, J. Mira, J. V. Sanchez-Andres (eds), Springer-Verlag, LNCS 1607, pp.341-249, ISBN: 3-540-66068-2

- Real time optical flow computation based on adaptive color quantization by competitive neural networks, M. Graña, I. Echave, Photonics East 99, Boston, USA, 19-22 Sept. 1999, en Intelligent Robots and Computer Vision XVIII pp. 165-174, D.P.Casasent (ed.) SPIE Press ISBN: 0-8194-3430-2

- VQ based image filtering, M. Graña, A.I. Gonzalez, I. Echave, J. Ruiz-Cabello, en Pattern Recognition and Image Analysis pp.471-478 M.I. Torres, A. Sanfeliu (eds.) ISBN: 84-95120-80-1

- VQ based Bayesian image filtering, M. Graña, I. Echave, J. Ruiz-Cabello, ICIP2000, IEEE press, ISBN: 0-7803-6300-0

- Neural network segmentation and classification of serial T1-weighted imaging of murine Aspergillus fumigatus thigh myositis, P. Aviles, I. Rodriguez, I. Echave, M. Graña, J. Ruiz-Cabello, D. Gargallo-Viola; ATLA, Third world Congress on Alternatives and Animal Use in the Life Sciences, Bologna, Italy, 29 Agosto-2 Septiembre 1999

- Increased robustness in visual processing with SOM-based filtering, E. Fernandez, I. Echave, M. Graña, IJCNN'2000, IEEE, AI*IA, Como, Italia, Julio 2000, S-I Amari, C. Lee Giles, M. Gori, V. Piuri (eds), IEEE press, pp. VI-131-134, ISBN: 0-7695-0619-4

- Competitive Neural Networks for robust computation of optical flow, E. Fernandez, I. Echave, M. Graña, ESANN'2000, Brujas, Be, Abril 2000, en ESANN 2000, M. Verleysen (ed.) pp. 89-94 Brussels: D-Facto publications ISBN: 2-930307-00-5

- Magnetic resonance imaging to monitor the biology of inflammation, J. Ruiz Cabello, I Rodriguez, P. Avilés, M. Graña, I. Echave, J. Regadera, D. Gargallo, M. Cortijo, 3th European Biophysics Congress, Munchen Ge. Sept. 2000, en European Biophysics Journal (springer) 29(4-5) pp.363

- MR image processing using neural networks and vector quantization, I. Rodriguez, M. Graña, I. Echave, P. Barreiro, M. Cortijo, J. Ruiz-Cabello, ESMRMB'00, 7th Annual Meeting of the European Association ofMagnetic Resonance in Medicine and Biology, Paris, sept. 2000, en Magnetic Resonance Materials in Physics, Biology and Medicine (MAGMA) (elsevier) vol 11 supl 1, sept 2000, p.143

- Segmentation of infected tissues in MRI based on VQ-BF filtering, M. Graña, I. Echave, J. Ruiz-Cabello, M. Cortijo, ICSP'02, Beijing, China, Agosto 2002, en ICSP'02 pp. 1540-1543, Yuan B., Tang X. (eds.) IEEE Press, ISBN: 0-7803-7488-6

Las siguientes publicaciones en revistas se produjeron también durante el desarrollo de la tesis:

- Monitoring acute inflammatory processes in the mouse muscle by MR imaging and spectroscopy: a comparision with pathological results, J. Ruiz-Cabello, M. Cortijo, I. Echave, M. Graña y otros, NMR in Biomedicine (2002) 15:204-214

- Computer-assisted enhanced volumetric segmentation magnetic resonance imaging data using a mixture of artificial neural networks, R. Perez de Alejo, J. Ruiz-Cabello, M. Cortijo, I. Echave, J. Regadera, J. Arrazola, P. Avilés, P. Barreiro, D. Gargallo, M. Graña, Magnetic Resonance Imaging (2003) 21(8):901-912

2. Redes neuronales competitivas

Las redes neuronales competitivas son la herramienta computacional fundamental en esta tesis. Este tipo de redes neuronales son en esencia métodos adaptativos para resolver problemas de agrupamiento (*Clustering*) y cuantización vectorial (*Vector Quantization*). Dado que los métodos de agrupamiento se pueden enmarcar en la teoría bayesiana de la decisión, en este capitulo comenzamos revisando los fundamentos y el entronque de estos procesos en dicho marco teórico.

En la sección 2.1 damos introducción al capítulo. En la sección 2.2 recordamos la formulación bayesiana del problema de la clasificación y su aplicación a las redes neuronales competitivas. En la sección 2.3 recordamos las diferencias entre la clasificación supervisada y la no supervisada. En la sección 2.4 recordamos el planteamiento del problema de entrenamiento no supervisado como un problema de estimación de máxima verosimilitud. En la sección 2.5 se plantea el problema sobre datos estacionarios. En la sección 2.6 se presentan las arquitecturas de redes neuronales competitivas consideradas. En la sección 2.7 se plantea el problema de la cuantización vectorial adaptativa sobre datos no estacionarios. En la sección 2.8 se presentan alguna conclusiones del capítulo.

2.1.Introducción

El agrupamiento particional basado en representantes de los agrupamientos y la cuantización vectorial son problemas relacionados que pueden ser mostrados como idénticos en términos generales. El agrupamiento (*Clustering*) [28] [39] [67] [135] es un proceso que trata de descubrir la estructura en la muestra de datos. La calidad de la solución se mide por una función criterio, basada en alguna medidad de similitud, que debe incorporar todos los aspectos que definen un buen agrupamiento para una aplicación dada.

La cuantización vectorial (*Vector Quantization*) [42] trata de encontrar un conjunto de representantes que minimice el error de cuantización esperado, basado en alguna distancia apropiada. Ambos problemas se pueden formular como proble-

mas de optimización no convexos y no lineales. Las redes neuronales competitivas son una clase de procedimientos de minimización del gradiente estocástico que se han venido aplicando a estos problemas.

La regla de aprendizaje competitivo básica, que denominamos Aprendizaje Competitivo Simple (*Simple Competitive Learning*) (SCL), se puede deducir como un algoritmo de minimización de la varianza intra-grupo en terminología de agrupamiento o, en terminología de cuantización vectorial, de la distorsión de la cuantización, mediante el método de descenso de gradiente estocástico. En teoría, otras reglas competitivas pueden derivarse como la minimización mediante el gradiente estocástico de otras funciones objetivo. Sin embargo, la mayor parte de las reglas de aprendizaje de redes neuronales competitivas han sido propuestas sobre un razonamiento intuitivo y es, en algunos casos, extremadamente difícil o imposible deducir analíticamente la formulación de la función objetivo que sería minimizada por estas reglas.

Para estudiar la convergencia de las distintas reglas competitivas asumimos que la aplicación final es el diseño de un cuantizador vectorial minimizando una función objetivo que es el error cuadrático medio (*least mean squares*) o distorsión euclídea. En otras palabras, estamos considerando que todas las posibles funciones objetivo se convierten, cuando se lleva al límite un parámetro de control, en la distorsión. La regla de aprendizaje básica y que debe ser superada por proposiciones más sofisticadas es el algoritmo SCL. Vemos a las otras reglas competitivas como elaboraciones sobre el SCL que convergen a él en función del valor de un parámetro de control. Las motivaciones, bajo este punto de vista, para la proposición de estas reglas competitivas serían dos:

1. Mejorar la velocidad de convergencia del algoritmo, o equivalentemente, mejorar la calidad de los resultados con el mismo costo computacional [23].

2. Obtener robustez frente a las variaciones en las condiciones iniciales.

Las arquitecturas concretas que vamos a considerar son el mapa auto-organizativo (*Self Organizing Map*) (SOM) [79], [80], una versión online del algoritmo de cuantización vectorial borrosa (*Fuzzy Learning Vector Quantization*) (FLVQ)[1] [8] y

[1]Una nota de precaución: la proposición original de FLVQ ponía el énfasis en su aplicacición como un algoritmo batch. Aplicamos esta definición como un algoritmo online porque tratamos de ajustar el algoritmo al esquema de entrenamiento en un solo paso sobre la muestra. Sin embargo, para mostrar los resultados completos hemos incluido resultados experimentales de las versiones batch y online del algoritmo FLVQ.

el algoritmo de competición suave (*Soft-Competition Scheme*) (SCS) [8], [153]. Las distintas arquitecturas de entrenamiento competitivo se caracterizan por sus funciones de vecindad respectivas [2]. Los parámetros de las funciones de vecindad controlan la expresión precisa de la función objetivo minimizada por la regla de aprendizaje competitivo y su convergencia hacia el SCL. Nos hemos concentrado en SOM, FLVQ y SCS. La programación eficiente de estos parámetros de control puede proporcionar algoritmos rápidos, robustos y eficientes para la minimización de la distorsión.

El problema de la generalización de los resultados y conclusiones a funciones de distorsión generalizadas radica en la definición apropiada de la secuencia de funciones, las reglas adaptativas y el control de la convergencia funcional.

La principal resticción computacional en los experimentos que se describen es la adaptación en un solo paso sobre la muestra. Hemos impuesto también la misma programación de la velocidad de aprendizaje en todos los algoritmos para tratar de aislar los efectos y la influencia de las distintas funciones de vecindad. Por el mismo motivo, la reducción de los vecindarios es la misma función exponencial para SOM, FLVQ, SCS. Los mejores resultados se obtienen cuando la la convergencia funcional al SCL es rápida.

Otras aproximaciones a la mejora de la robustez del SCL encontradas en la literatura, como las basadas en el mecanismo de conciencia [2] [27] o el Neural-Gas [96] no han sido considerados en este trabajo porque no encontramos que sea inmediata la convergencia funcional de sus funciones objetivo a la distorsión euclídea.

La consideración de datos variantes en el tiempo conduce a la formulación de la cuantización vectorial adaptativa y el agrupamiento no estacionario. Formulamos estos problemas como la búsqueda dinámica de los representantes óptimos basandose en los datos muestreados en cada instante de tiempo. Las redes neuronales competitivas han sido de poca utilidad en este problema debido a su lenta convergencia. [42]. Las aproximaciones más comunes están basadas en los algoritmos de relleno de los libros de código [35], los cuales plantean serios problemas de ajuste. Las redes neuronales con entrenamiento en un paso tienen las siguientes características que las hacen apropiadas para este problema:

1. Se pueden afinar fácilmente (especialmente el SOM).

2. Son muy robustas y son relativamente rápidas.

[2] Que son denominadas funciones de interferencia en [72]

3. Su complejidad computacional crece linealmente con la dimensión del espacio y los tamaños de los libros de código y la muestra.

2.2.Formulación bayesiana de los problemas de clasificación

Comenzamos introduciendo la notación básica que vamos a utilizar:

- $\mathbf{x} \in \mathbb{R}^d$ vector aleatorio de características que se calcula sobre los objetos en el mundo.
- $\Omega = \{\omega_1, .., \omega_c\}$ conjunto finito de estados de la naturaleza o clases.
- $A = \{\alpha_1, .., \alpha_a\}$ conjunto de acciones posibles resultado de una decisión.
- $\lambda(\alpha_i | \omega_j)$ costo de realizar la acción α_i cuando se detecta que el estado de la naturaleza es ω_j.
- $p(\mathbf{x} | \omega_j)$ densidad de probabilidad (d.d.p.) de $\mathbf{x}$ condicionada al estado de la naturaleza o clase, también denominada *verosimilitud*.
- $p(\mathbf{x})$ densidad de probabilidad total de $\mathbf{x}$ independiente del estado de la naturaleza, también denominada *evidencia*.
- $P(\omega_j | \mathbf{x})$ probabilidad *a posteriori* del estado de la naturaleza dado el vector de características $\mathbf{x}$.
- $P(\omega_j)$ probabilidad *a priori* del estado de la naturaleza.

Definition 1. ***Riesgo condicional****: esperanza de la pérdida (costo) asociada a una acción*

$$R(\alpha_i | \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i | \omega_j) P(\omega_j | \mathbf{x}) \tag{2.1}$$

Definition 2. ***Riesgo total****: esperanza de la pérdida extendida al espacio de características*

$$R = \int R(\alpha(\mathbf{x}) | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{2.2}$$

donde $\alpha(\mathbf{x})$ *denota la regla de decisión* $\alpha : \mathbb{R}^d \to A$

Definition 3. *La* ***regla Bayesiana de decisión*** *minimiza el riesgo total R: escoge α_i para la que $R(\alpha_i|\mathbf{x})$ es mínimo.*

Estas tres definiciones establecen que cualquier algoritmo de decisión puede definirse como un algoritmo de minimización de una función objetivo, el riesgo total, y que la estrategia óptima es la bayesiana por que garantiza la minimización de dicha función objetivo. La clasificación puede considerarse como un caso especial de algoritmo de decisión, si consideramos que las acciones consisten en el reconocimiento del estado de la naturaleza

$$\alpha_i \equiv \omega_i \quad i = 1, .., c. \tag{2.3}$$

La función objetivo es la función de pérdida simétrica $0-1$ que se define formalmente como

$$\lambda_{i,j} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, .., c. \tag{2.4}$$

El riesgo condicional toma la forma del error de clasificación, esto es la probabilidad de que el objeto perteneza a otra clase distinta de aquella a la que lo hemos asignado:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1;j\neq i}^{c} P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x}) \tag{2.5}$$

La regla bayesiana de decisión se convierte en la regla de clasificación de error mínimo y se puede fomular como sigue:

$$\text{Decide } \omega = \omega_j \text{ si } P(\omega_j|\mathbf{x}) > P(\omega_k|\mathbf{x}); \forall k \neq j \tag{2.6}$$

En general se considera un conjunto de funciones asociadas a las clases, las funciones discriminantes

$$g_i(\mathbf{x}) \quad i = 1, .., c \tag{2.7}$$

de forma que el sistema que implementa la regla de clasificación puede expresarse de forma canónica como

$$\text{Decide } \omega = \omega_i \text{ si } i = \underset{k=1..c}{argmax} \{g_k(\mathbf{x})\}. \tag{2.8}$$

Usando estas funciones discriminantes, se definen las regiones de decisión

$$\mathcal{R}_j = \left\{ \mathbf{x} \left| j = \underset{k=1..c}{argmax} \{g_k(\mathbf{x})\} \right. \right\}; j = 1..c \tag{2.9}$$

que particionan el espacio de características en función de la clasificación asignada a los vectores de características por la regla de clasificación.

Para el clasificador Bayesiano de mínima razón de error la función discriminante es directamente la probabilidad *a posteriori* de la clase:

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)} \tag{2.10}$$

que puede reescribirse sin que varíen las regiones de decision resultantes de las siguientes maneras

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i), \tag{2.11}$$

$$g_i(\mathbf{x}) = log\ p(\mathbf{x}|\omega_i) + log\ P(\omega_i) \tag{2.12}$$

Para dos regiones de decisión contiguas, la ecuación

$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \tag{2.13}$$

define la superficie de decisión que las separa.

2.2.1.Funciones discriminantes basadas en la distribución normal

La función de densidad de probabilidad normal univariante es

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

donde $E[x] = \mu$ y $E\left[(x-\mu)^2\right] = \sigma^2$, donde $E[.]$ es la esperanza matemática. En el caso multivariante $\mathbf{x} \in \mathbb{R}^d$ la función de densidad de probabilidad normal tiene la siguiente expresión:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

donde $E[\mathbf{x}] = \boldsymbol{\mu} = [\mu_i]$ y $E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\right] = \Sigma = [\sigma_{ij}]$. La matriz de covarianza Σ es simétrica y semidefinida positiva. Habitualmente se asume que es definida positiva $|\mathbf{\Sigma}| > 0$. Como es sabido, si $\mathbf{\Sigma}$ es diagonal, los componentes del vector aleatorio $\mathbf{x}$ son independientes: $p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i)$. Los puntos de idéntica probabilidad tienen distancia de Mahalanobis $\mathbf{r}^2 = (\mathbf{x}-\boldsymbol{\mu})^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ constante y las superficies de equiprobabilidad tienen forma hiperelipsoidal.

Cuando las funciones de densidad de probabilidad condicional del vector de características tienen la forma normal el tipo de función discriminante natural es logarítmico

$$g_j(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) - \frac{1}{2} log\,|\boldsymbol{\Sigma}_j| + log P(\omega_j). \tag{2.14}$$

En el caso particular en que las matrices de covarianza son la identidad y las probabilidades a priori de las clases son idénticas, la función discriminante consiste en la distancia euclídea respecto del centro de la clase:

$$g_j(\mathbf{x}) = -\left\|\mathbf{x}-\boldsymbol{\mu}_j\right\| \tag{2.15}$$

2.3.Aprendizaje supervisado versus no supervisado

Dada una muestra $S = \{\mathbf{x}_1, .., \mathbf{x}_n\}$ de los vectores de características, usualemente se entiende por resolver el problema de la clasificación la construcción del clasificador, esto es la estimación de los parámetros de los modelos probabilísticos involucrados: probabilidades *a priori* de las clases, parámetros de las densidades condicionales, o la estimación directa de las probabilidades *a posteriori*.

El aprendizaje supervisado (o mejor la construcción supervisada del clasificador) asume que la muestra está particionada asignandose a cada vector de características la clase de la que ha sido extraido:

$$\begin{aligned} S &= S_1 \cup S_2 \cup ... \cup S_c, \\ S_i \cap S_j &= \emptyset; i \neq j. \end{aligned} \tag{2.16}$$

En esta situación algunos parámetros del clasificador, como las probabilidades *a priori* de las clases pueden estimarse de forma inmediata:

$$P(\omega_i) = \frac{|S_i|}{|S|}. \tag{2.17}$$

Otros parámetros, como los de los modelos de las funciones probabilidad condicionada asumidas, pueden estimarse maximizando la verosimilitud de la parte de la muestra correspondiente a una clase. Sea $S_i = \{\mathbf{x}_1^i, .., \mathbf{x}_n^i\}$ la partición de la muestra correspondiente a la clase ω_i, y $p_{\boldsymbol{\theta}}(\mathbf{x}\,|\omega_i)$ el modelo de la función de densidad, donde $\boldsymbol{\theta}$ es el vector de parámetros desconocidos. Si las muestras en

S_i provienen independientemente de $p_{\boldsymbol{\theta}}(\mathbf{x}|\omega_i)$, la probabilidad de obtener S es proporcional a

$$p_{\boldsymbol{\theta}}(S_i) = \prod_{k=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}_k|\omega_i)$$

donde $p_{\boldsymbol{\theta}}(S)$ es la verosimilitud de $\boldsymbol{\theta}$ respecto de S. Dado el modelo $p_{\boldsymbol{\theta}}(\mathbf{x})$ la estimación maximo verosimil de $\boldsymbol{\theta}$ respecto de S es el valor $\widehat{\boldsymbol{\theta}}$ que maximiza $p_{\boldsymbol{\theta}}(S)$ como función de $\boldsymbol{\theta}$. Usualmente se maximiza el logarítmo de la verosimilitud para simplificar los cálculos.

Por ejemplo, en el caso de que el modelo de la función de densidad sea la función normal $p_{\boldsymbol{\theta}}(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ los parámetros a estimar son $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}_i$, y la estimación maximo verosimil viene dada por

$$\widehat{\boldsymbol{\mu}}_i = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k^i, \tag{2.18}$$

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n}\sum_{k=1}^{n} \left(\mathbf{x}_k^i - \widehat{\boldsymbol{\mu}}_i\right)\left(\mathbf{x}_k^i - \widehat{\boldsymbol{\mu}}_i\right)^t. \tag{2.19}$$

Si no se dispone de la información sobre la clase a la que pertenece cada vector en la muestra, el aprendizaje es no supervisado. El problema de la construcción no supervisada del clasificador se convierte en un problema combinatorio, en el que deberiamos evaluar los parámetros del modelo probabilístico para cada posible partición de la muestra y escoger la partición de mayor verosimilitud. Obviamente, esta aproximación es inviable para muestras de un tamaño no pequeño. Por ello, todos los métodos que se proponen para el aprendizaje no supervisado son subóptimos en la medida en que son métodos de optimización locales que no resuelven el problema de optimización global sino que encuentran un óptimo local.

2.4.Aprendizaje no supervisado

Consideramos conocidos

1. el número de clases c
2. las probabilidades *a priori* $P(\omega_i)$ $i = 1, .., c$
3. los modelos de las d.d.p. condicionadas $p(\mathbf{x}|\omega_i; \boldsymbol{\theta}_i)$ $i = 1, .., c$

Los parámetros desconocidos se agregan en el vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_i; i = 1, .., c)$. Consideramos en general que los vectores de características siguen de una d.d.p. mezcla de las d.d.p. condicionadas, lo que se puede expresar de la forma

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{c} p(\mathbf{x}|\omega_i; \boldsymbol{\theta}_i) P(\omega_i), \tag{2.20}$$

donde las probabilidades *a priori* de las clases $P(\omega_i)$ son considerados como los parámetros de la mezcla y las $p(\mathbf{x}|\omega_i; \boldsymbol{\theta}_i)$ son las densidades componente. La d.d.p. de los vectores de características $p(\mathbf{x}|\boldsymbol{\theta})$ es identificable si para todo $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$ existe al menos un valor del vector de características $\mathbf{x}^*$ para el que $p(\mathbf{x}^*|\boldsymbol{\theta}') \neq p(\mathbf{x}^*|\boldsymbol{\theta})$. En el aprendizaje no supervisado si $p(\mathbf{x}|\boldsymbol{\theta})$ no es identificable la mejor solución que podemos obtener consiste en la partición del espacio de características en regiones que corresponden a las clases pero sin que podamos indentificar la clase precisa. Esto es, la mejor solución posible consiste en una permutación aleatoria de las clases de la solución óptima.

2.4.1.Estimación no supervisada de maxima verosimilitud

Consideramos $S = \{\mathbf{x}_1, .., \mathbf{x}_n\}$ muestras del vector de características que sigue la d.d.p. $p(\mathbf{x}|\boldsymbol{\theta})$. La verosimilitud de S es

$$p(S|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}) \tag{2.21}$$

y el estimador maximo verosimil es el valor del vector de parámetros $\widehat{\boldsymbol{\theta}}$ que maximiza $p(S|\boldsymbol{\theta})$. La log-verosimilitud

$$l = log\ p(S|\boldsymbol{\theta}) = \sum_{k=1}^{n} log\ p(\mathbf{x}_k|\boldsymbol{\theta}) \tag{2.22}$$

tiene los máximos en los mismo puntos que la función de verosimilitud, es equivalente a ella a efectos de estimación de los vectores de parámetros óptimos y mucho más manejable cuando las densidades de probabilidad son exponenciales. Asumimos que $p(S|\boldsymbol{\theta})$ es diferenciable respecto de $\boldsymbol{\theta}$. El gradiente de la log-verosimilitud l respecto del vector de parámetros desconocidos de un componente

de la mezcla es de la forma:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_i} l &= \sum_{k=1}^{n} \frac{\nabla_{\boldsymbol{\theta}_i} p\left(\mathbf{x}_k \left| \boldsymbol{\theta} \right.\right)}{p\left(\mathbf{x}_k \left| \boldsymbol{\theta} \right.\right)} \\ &= \sum_{k=1}^{n} \frac{\nabla_{\boldsymbol{\theta}_i} \sum_{j=1}^{c} p\left(\mathbf{x}_k \left| \omega_j; \boldsymbol{\theta}_j \right.\right) P\left(\omega_j\right)}{p\left(\mathbf{x}_k \left| \boldsymbol{\theta} \right.\right)}. \end{aligned} \tag{2.23}$$

Dado que $\boldsymbol{\theta}_i$ y $\boldsymbol{\theta}_j$ son funcionalmente independientes si $i \neq j$, y sustituyendo la expresión de la probabilidad *a posteriori* obtenemos

$$\nabla_{\boldsymbol{\theta}_i} l = \sum_{k=1}^{n} P\left(\omega_i \left| \mathbf{x}_k; \boldsymbol{\theta} \right.\right) \nabla_{\boldsymbol{\theta}_i} log\ p\left(\mathbf{x}_k \left| \omega_i; \boldsymbol{\theta} \right.\right). \tag{2.24}$$

Si la función de log-verosimilitud es convexa, el estimador de maxima verosimilitud es el $\widehat{\boldsymbol{\theta}}_i$ que anula el gradiente y puede obtenerse como solución del sistema de ecuaciones

$$\sum_{k=1}^{n} P\left(\omega_i \left| \mathbf{x}_k;\ \widehat{\boldsymbol{\theta}}_i \right.\right) \nabla_{\boldsymbol{\theta}_i} log\ p\left(\mathbf{x}_k \left| \omega_i; \widehat{\boldsymbol{\theta}} \right.\right) = 0 \quad i = 1, .., c \tag{2.25}$$

En el caso habitual en que desconocemos las probabilidades *a priori* de las clases, se deducen de la condición de gradiente las siguientes expresiones para los estimadores de máxima verosimilitud:

$$\widehat{P}\left(\omega_i\right) = \frac{1}{n} \sum_{k=1}^{n} \widehat{P}\left(\omega_i \left| \mathbf{x}_k; \widehat{\boldsymbol{\theta}}_i \right.\right) \quad i = 1, .., c, \tag{2.26}$$

$$\sum_{k=1}^{n} \widehat{P}\left(\omega_i \left| \mathbf{x}_k; \widehat{\boldsymbol{\theta}}_i \right.\right) \nabla_{\boldsymbol{\theta}_i} log\ p\left(\mathbf{x}_k \left| \omega_i; \widehat{\boldsymbol{\theta}} \right.\right) = 0 \quad i = 1, .., c, \tag{2.27}$$

$$\widehat{P}\left(\omega_i \left| \mathbf{x}_k; \widehat{\boldsymbol{\theta}}_i \right.\right) = \frac{p\left(\mathbf{x}_k \left| \omega_i; \widehat{\boldsymbol{\theta}}_i \right.\right) \widehat{P}\left(\omega_i\right)}{\sum_{J=1}^{c} p\left(\mathbf{x}_k \left| \omega_j; \widehat{\boldsymbol{\theta}}_i \right.\right) \widehat{P}\left(\omega_j\right)}. \tag{2.28}$$

La ecuación 2.28 proporciona una aproximación a la partición de la muestra en muestras de cada clase. La probabilidad *a posteriori* puede considerarse como un coeficiente suave de pertenencia del elemento de la muestra a la clase. De esta forma se resuelve simultaneamente el problema combinatorio de la asignación de

muestras a clases y el de la estimación de los modelos probabilísticos que se basan en las ecuaciones 2.26 y 2.27. La iteración de las ecuaciones 2.26-2.28 a partir de estimaciónes iniciales arbitrarias es un método de optimización local que sería valido para funciones con un único máximo, lo que obviamente no es el caso de la función de log-verosimilitud. Este conjunto de ecuaciones se corresponde con el conocido esquema de Maximización de la Expectación (Expectation Maximization) en el que las variables ocultas son las probabilidades *a posteriori* de las clases. El paso E lo constituyen las ecuaciones 2.26 y 2.27 mientras que el paso M lo constituye la ecuación 2.28.

Cuando las densidades componente de la mezcla son densidades normales:

$$p\left(\mathbf{x}\left|\omega_i;\boldsymbol{\theta}_i\right.\right) \sim N\left(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right) \tag{2.29}$$

y consideramos que los parámetros desconocidos son sólo las medias $\boldsymbol{\mu}_i \equiv \boldsymbol{\theta}_i$, estamos ante un problema de agrupamiento basado en representantes, o de cuantización vectorial. Si sustituimos la expresión de la d.d.p. normal en las ecuaciones 2.26-2.28 obtenemos las ecuaciones que aplicadas de forma iterativa a partir de buenas estimaciónes iniciales $\boldsymbol{\mu}_i\left(0\right)$ de las medias de las clases, nos permiten calcular las estimaciónes de las medias y de la partición de la muestra:

$$\begin{aligned} \widehat{\boldsymbol{\mu}}_i &= \sum_{k=1}^{n} \mathbf{x}_k \frac{\widehat{P}\left(\omega_i\left|\mathbf{x}_k;\widehat{\boldsymbol{\mu}}\right.\right)}{\sum_{k=1}^{n}\widehat{P}\left(\omega_i\left|\mathbf{x}_k;\widehat{\boldsymbol{\mu}}\right.\right)}; i=1,...,c, \\ \widehat{P}\left(\omega_i\left|\mathbf{x}_k;\widehat{\boldsymbol{\mu}}\right.\right) &= \frac{p\left(\mathbf{x}_k\left|\omega_i;\widehat{\boldsymbol{\mu}}\right.\right)P\left(\omega_i\right)}{\sum_{j=1}^{c}p\left(\mathbf{x}_k\left|\omega_j;\widehat{\boldsymbol{\mu}}\right.\right)P\left(\omega_j\right)} \end{aligned} \tag{2.30}$$

Cuando se asume que las matrices de covarianza son la identidad $\boldsymbol{\Sigma}_i = \mathbf{I}$ para todas las clases y se aplica como regla de decisión la regla bayesiana $g_i\left(\mathbf{x}_k\right) = -\widehat{P}\left(\omega_i\left|\mathbf{x}_k;\widehat{\boldsymbol{\mu}}\right.\right)$ para determinar la asignación de un vector de la muestra a una clase y realizar la partición especificada por la ecuación 2.16:

$$\mathbf{x}_k \in S_i \Leftrightarrow i = \underset{j=1..c}{argmin}\left\{\left\|\mathbf{x}_k - \boldsymbol{\mu}_j\right\|\right\}, \tag{2.31}$$

entonces tenemos el algoritmo más simple de aprendizaje no supervisado. Este algoritmo ha sido denominado de muy diversas maneras: Isodata simple [28], K-medias [39], LBG en tratamiento de señal. Este algoritmo consiste en, a partir de unas estimaciónes iniciales arbitrarias $\boldsymbol{\mu}_1\left(0\right),..,\boldsymbol{\mu}_c\left(0\right)$, iterar la estimación de las medias de las clases

$$\boldsymbol{\mu}_i\left(t+1\right) = \sum_{\mathbf{x}\in S_i}\frac{\mathbf{x}}{\left|S_i\right|}, \tag{2.32}$$

donde la partición de la muestra se ha calculado de acuerdo a la ecuación 2.31. Es importante notar que este algoritmo se deduce como la maximización de la función de log-verosimilitud, que bajo las restricciones que llevan al algoritmo de K-medias, es equivalente a la minimización de la función de distorsión euclídea de la cuantización de la muestra usando los representantes de las clases como vector de códigos:

$$\xi_E^2 = \sum_{i=1}^{c} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{2.33}$$

Se puede ampliar el algoritmo incluyendo las estimaciónes de las matrices de covarianza, pero no entraremos en estos detalles aquí. Hemos de notar que el algoritmo de K-medias corresponde a un método de agrupamiento en el que la distancia entre vectores es la euclídea y la función objetivo es la distorsión euclídea.

2.5. Agrupamiento y cuantización vectorial sobre datos estacionarios.

Antes de presentar las redes neuronales competitivas, revisaremos la definición de los problemas genéricos a los que se aplican, los de la cuantización vectorial (*Vector Quantization*) y el agrupamiento (*Clustering*) basado en particiones. Seguimos el trabajo clásico de [42] con variaciones pequeñas de notación.

Un cuantizador vectorial $\mathcal{Q}$ de dimensión d y tamaño c es un mapa de un espacio real d-dimensional en un conjunto finito $\mathbf{Y}$ de vectores representantes (*codevectors*), usualmente denominado el libro de códigos (*codebook*).

$$\mathcal{Q} : \mathbb{R}^d \to \mathbf{Y} \tag{2.34}$$

donde $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_c\}$ y cada $\mathbf{y}_i \in \mathbb{R}^d$ con $i \in \mathcal{I} = \{1, 2, ..., c\}$. Asociado con cada libro de códigos hay una partición del espacio input:

$$R_i = \left\{\mathbf{x} \in \mathbb{R}^d \,|\, \mathcal{Q}(\mathbf{x}) = \mathbf{y}_i\right\} \tag{2.35}$$

$$\bigcup_{i=1}^{c} R_i = \mathbb{R}^d; \bigcap_{i=1}^{c} R_i = \emptyset \tag{2.36}$$

Un cuantificador vectorial se puede descomponer en dos operaciones : el codificador vectorial $\mathcal{E}$ y el decodificador vectorial $\mathcal{D}$.

$$\mathcal{Q}(\mathbf{x}) = \mathcal{D} \cdot \mathcal{E}(\mathbf{x}) = \mathcal{D}(\mathcal{E}(\mathbf{x})) = \widehat{\mathbf{x}}. \tag{2.37}$$

El codificador realiza la correspondencia del vector input en el conjunto índice, el decodificador realiza realiza la correspondencia inversa, desde el conjunto índice al espacio input devolviendo el vector código indexado. :

$$\begin{aligned} \mathcal{E} &: \mathbb{R}^d \to \mathcal{I} \\ \mathcal{D} &: \mathcal{I} \to \mathbb{R}^d \end{aligned} \tag{2.38}$$

El codificador está completamente determinado por la partición definida por la expresión 2.35, mientras que el decodificador está determinado por el libro de códigos. Esta notación está claramente sesgada hacia las aplicaciones de compresión de señal, donde el codificador produce el código a ser enviado o almacenado y el decodificador recupera una aproximación a la señal original. Dada una medida del error de reproducción

$$\varepsilon\left(\mathbf{x},\widehat{\mathbf{x}}\right) = \varepsilon\left(\mathbf{x},\mathcal{D}\left(\mathcal{E}\left(\mathbf{x}\right)\right)\right), \tag{2.39}$$

el rendimiento del cuantizador vectorial se mide por la esperanza de este error de reproducción

$$\xi = E_{\mathbf{x}}\left[\varepsilon\left(\mathbf{x},\widehat{\mathbf{x}}\right)\right] = \int \varepsilon\left(\mathbf{x},\widehat{\mathbf{x}}\right) p\left(\mathbf{x}\right) d\mathbf{x}. \tag{2.40}$$

donde $p\left(\mathbf{x}\right)$ es la d.d.p. de los vectores input. Cuando la fuente que genera los vectores input es estacionaria esta d.d.p. permanece invariante en el tiempo. Dada una muestra $\aleph = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ el rendimiento definido por la expresión 2.40 puede ser estimado calculando el error medio sobre la muestra

$$\widehat{\xi} = \frac{1}{n}\sum_{j=1}^{n} \varepsilon\left(\mathbf{x}_j,\widehat{\mathbf{x}}_j\right). \tag{2.41}$$

Teniendo en cuenta la estructura del cuantizador vectorial, podemos reescribir estas expresiones como sigue (P_i es la probabilidad *a priori* de la región):

$$\begin{aligned} \xi &= \sum_{i=1}^{c} P_i \int_{R_i} \varepsilon\left(\mathbf{x},\mathbf{y}_i\right) p\left(\mathbf{x}\,|\mathbf{x} \in R_i\right) d\mathbf{x} \\ &= \sum_{i=1}^{c} P_i E_{\mathbf{x}}\left[\varepsilon\left(\mathbf{x},\mathbf{y}_i\right)|\mathbf{x} \in R_i\right]. \end{aligned} \tag{2.42}$$

y su estimación se puede reescribir como sigue

$$\widehat{\xi} = \frac{1}{n}\sum_{i=1}^{c}\sum_{\mathbf{x}_j \in R_i} \varepsilon\left(\mathbf{x}_j,\mathbf{y}_i\right). \tag{2.43}$$

La medida de error más extensamente usada, y la que vamos a usar efectivamente en nuestros trabajos experimentales es la distancia euclídea

$$\varepsilon_E^2(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2 = (\mathbf{x}-\mathbf{y})^t(\mathbf{x}-\mathbf{y}), \tag{2.44}$$

que es la base para el cálculo de la distorsión euclídea

$$\xi_E^2 = \sum_{i=1}^{c} P_i E_{\mathbf{x}}\left[\|\mathbf{x}-\mathbf{y}_i\|^2 \,|\mathbf{x}\in R_i\right] \tag{2.45}$$

$$\widehat{\xi}_E^2 = \sum_{i=1}^{c}\frac{1}{c}\sum_{j=1}^{n}\|\mathbf{x}-\mathbf{y}_i\|^2\,\delta_{ij} \tag{2.46}$$

Una clase especial de cuantizadores vectoriales son los cuantizadores del vecino más cercano, en el que la partición del espacio input se define como sigue

$$R_i = \{\mathbf{x}\,|i = argmin\,\{\varepsilon(\mathbf{x},\mathbf{y}_i); i=1,...,c\}\}, \tag{2.47}$$

cuando la medida de similitud es la distancia euclídea al cuadrado, los cuantizadores del vecino más cercano se convierten en cuantizadores de Voronoi, basados en una teselación de Voronoi del espacio input. Debe notarse que la partición del espacio input inducida por el cuantizador del vecino más cercano depende fuertemente de la medida de error considerada.

Definition 4. *El problema del diseño de un cuantizador vectorial es la búsqueda del libro de códigos que minimiza la distorsión*

$$\min_{\mathbf{Y}}\ \xi \tag{2.48}$$

Basado en la muestra de datos, este problema se formula como la minimización de la distorsión estimada sobre la muestra

$$\min_{\mathbf{Y}}\ \widehat{\xi} \tag{2.49}$$

El agrupamiento particional [28], [67], [135], [39] es un método estadístico no paramétrico de exploración de los datos. Es la búsqueda de una partición de los datos que optimiza una función de criterio de agrupamiento dada. En la notación introducida arriba:

$$\min_{\{\aleph_1,...,\aleph_c\}}\ \widehat{\xi} \tag{2.50}$$

donde $\{\aleph_1, ..., \aleph_c\}$ es la partición de la muestra inducida por la partición del espacio $\{R_1, ..., R_c\}$. Hacemos énfasis en la naturaleza estacionaria de esta definición para distinguirla claramente del agrupamiento en el caso no estacionario que se introduce más adelante. El problema del diseño del cuantizador vectorial basado en los datos de la muestra es equivalente al del agrupamiento particional de la muestra cuando el criterio de agrupamiento es una generalización de la varianza entre agrupamientos. Cuando la medida de error de la cuantización es la distancia euclídea al cuadrado y la partición se basa en la regla del vecino más cercano, tanto la cuantización vectorial como el agrupamiento se pueden formalizar como el siguiente problema de optimización

$$\min_{\mathbf{Y}} \widehat{\xi}_E^2 = \min_{\mathbf{Y}} \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} \|\mathbf{x}_j - \mathbf{y}_i\|^2 \delta_{ij} \tag{2.51}$$

donde δ_{ij} el coeficiente de pertenencia es la siguiente función

$$\delta_{ij} = \begin{cases} 1 & i = \arg\min\limits_{k=1,..,c} \left\{ \|\mathbf{x}_j - \mathbf{y}_k\|^2 \right\}, \\ 0 & \text{sino.} \end{cases} \tag{2.52}$$

Se han propuesto un gran número de métodos para resolver este problema. Para los cuantizadores basados en la regla del vecino más cercano el más popular es el algoritmo de Lloyds generalizado (*Generalized Lloyds Algorithm*) (GLA) [42] también denominado como Isodata en la literatura de reconocimiento de patrones o K-medias en análisis estadístico de datos [28] [39] [135]. Dado un libro de códigos inicial $\mathbf{Y}(0) = (\mathbf{y}_1(0), \mathbf{y}_2(0), ..., \mathbf{y}_c(0))$ el algoritmo de K-medias realiza la iteración

$$\mathbf{y}_i(t) = \frac{\sum_{j=1}^{n} \mathbf{x}_j \delta_{ij}(t-1)}{\sum_{j=1}^{n} \delta_{ij}(t-1)}; i = 1, .., c; t = 1, 2, ... \tag{2.53}$$

hasta que se produce una condición de parada, que puede ser definida en términos absolutos o relativos expresados por las ecuaciones 2.54 y 2.55, respectivamente

$$\widehat{\xi}_E^2(t) - \widehat{\xi}_E^2(t-1) < \epsilon, \tag{2.54}$$

$$\frac{\widehat{\xi}_E^2(t) - \widehat{\xi}_E^2(t-1)}{\widehat{\xi}_E^2(t-1)} < \epsilon. \tag{2.55}$$

El parámetro t es el número de iteración y $\delta_{ij}(t-1)$ es es la pertenencia evaluada usando el libro de códigos calculado en la iteración anterior. El algoritmo

K-medias o Isodata es un algoritmo de búsqueda local que da buenos resultados para buenascondiciones iniciales. Los intentos de combinar Isodata con el enfriamiento simulado para convertirlo en un procedimiento de minimización global son precedentes del esquema de competición suave (SCS) [138], [155] [120], [121], [122], [123].

Las fuentes del método de cuantificación vectorial borrosa (Fuzzy Learning Vector Quantization) (FLVQ) son los trabajos de Bezdek [6], quien propone un método de agrupamiento borroso basado en la minimización de la función criterio

$$J_m\left(U, \mathbf{Y}, \aleph\right) = \sum_{j=1}^{n} \sum_{i=1}^{c} \left(u_{ij}\right)^m \left\|\mathbf{x}_j - \mathbf{y}_i\right\|_A^2, \tag{2.56}$$

donde $\|.\|_A$ es una norma de producto interno que constituye una generalización lineal de la distancia euclídea. La solución óptima cuando $m > 1$ conduce a la condición de pertenencia borrosa [6] dada por

$$u_{ij} = \left(\sum_{k=1}^{c} \left(\frac{\left\|\mathbf{x}_j - \mathbf{y}_i\right\|_A^2}{\left\|\mathbf{x}_j - \mathbf{y}_k\right\|_A^2} \right)^{\frac{1}{m-1}} \right)^{-1}, \tag{2.57}$$

y a la condición de los centroides borrosos

$$\mathbf{y}_i = \frac{\sum_{j=1}^{n} \mathbf{x}_j \left(u_{ij}\right)^m}{\sum_{j=1}^{n} \left(u_{ij}\right)^m}, \tag{2.58}$$

como la extensión natural de las condiciones para la solución óptima del problema de agrupamiento duro basado en la minimización de la distorsión euclídea [42]. El algoritmo borroso de las c-medias (*Fuzzy c-means*) (FCM) es una reformulación del Isodata basado en las condiciones dadas por las ecuaciones 2.58 y 2.57. La pertenencia borrosa se convierte en la pertenencia dura en el límite del parámetro m,

$$\lim_{m \to 1^+} \left(u_{ij}\right)^m = \delta_{ij} \tag{2.59}$$

por lo que el FCM se convierte en el Isodata, y la función criterio del agrupamiento borroso se convierte en la distorsión (si A es la matriz identidad)

$$\lim_{m \to 1^+} J_m\left(U, \mathbf{Y}, \aleph\right) = \widehat{\xi_E^2} \tag{2.60}$$

2.5.1.Comentarios sobre el problema de la validación

Un punto importante es la validación de los algoritmos de agrupamiento y cuantización vectorial. Algunas veces, puede formularse como la optimización de una cierta función objetivo, que puede ser asumida como una generalización de la distorsión. Si varios algoritmos van a ser comparados en base a una función de distorsión, una comparación justa debe asegurarnos que cada uno de ellos está minimizando la misma o una función equivalente en el sentido de que los óptimos buscados caen en los mismos lugares del espacio de búsqueda. Este no es habitualmente el caso. Por esta razón, los autores recurren a procedimientos de validación indirectos. Por ejemplo [8], [72], [136], [108] usan los datos de las flores Iris de la siguiente manera:

1. Se aplica el algoritmo no supervisado de agrupamiento a los datos. Los agrupamientos encontrados se etiquetan.

2. Los algoritmos en competencia se comparan en base al número de errores de clasificación producidos en relación al etiquetado original de los datos Iris.

Pensamos que este procedimiento no da una verificación general de la validez y eficiencia de los algoritmo, entre otras cosas por que a veces no se tienen en cuenta la no identificabilidad de las clases reales mediante métodos no supervisados. Otras aproximaciones descansan en la evaluación subjetiva de los resultados. Por ejemplo en [95] el resultado final es la segmentación de las imágenes producida por el agrupamiento de las características de textura. En muchos casos de segmentación de imágenes basada en el agrupamiento [116], [109], [137], [143] los autores recurren a los resultados visuales como la validación final del procedimiento. También en [95] se usa un test de Kolmogorov-Smirnov sobre las distancias regularizadas de Mahalanobis entre representantes de agrupamientos bajo la asunción de agrupamientos gausianos para evaluar la compacidad de los agrupamientos resultantes. En [139], [140], [141] el objetivo es obtener cuantizaciones equiprobables del espacio input. Sus resultados son las probabilidades de activación de las unidades sobre conjuntos de datos específicos, y algunos ejemplos gráficos de la cuantización obtenida en espacios y d.d.p. específicos. Cuando el objetivo es la preservación topológica [79], [133] los resultados son representaciones gráficas de instancias cuantizadas. Para aplicaciones de codificación y comunicaciones, la distorsión euclídea y la tasa de señal-ruido (*Signal to Noise Ratio*)(SNR) se usan habitualmente como las medidas de rendimiento [42].

Por simplicidad y generalidad, hemos asumido que la distorsión es la medida de rendimiento apropiada para validar los algoritmos. Es la asunción mínima y la tomada en muchas instancias de la vida real, incluyendo el problema de la cuantización del color.

2.6.Redes Neuronales Competitivas

En el marco de las aplicaciones de agrupamiento y cuantización vectorial, la función de error está relacionada con la calidad del agrupamiento y con el error de cuantización, respectivamente. El algoritmo de minimización del gradiente estocástico de la función criterio del agrupamiento y de la distorsión euclídea, respectivamente, produce los distintos algoritmos de aprendizaje de las Redes Neuronales Competitivas (RNC) [62], [73], [79], [60], [135], [84].

Comenzamos con la definición del algoritmo de Aprendizaje Competitivo Simple (SCL) que minimiza la distorsión euclídea. Tras ello damos las formulaciones de los algoritmos SOM, FLVQ y SCS en el marco de la regla competitiva general. Discutimos la función objetivo minimizada por cada una estas reglas de aprendizaje y su relación con la distorsión euclídea. El argumento es que, como se discutía en [23], el SOM y otras RNC pueden ser tomadas como inicializaciones robustas para el SCL, cuando el objetivo es la minimización de la distorsión euclídea. Argumentos similares aparecen en la literatura [122], [123], [8], [136], [138], [153], [155] justificando los diversos intentos de definición de algoritmos robustos de agrupamiento o VQ.

2.6.1.Algoritmo Competitivo Simple

Resulta de la aplicación del método del gradiente estocástico a la minimización de la función objetivo dada por la distorsión de la muestra respecto de los centros de las clases usando la distancia euclídea [39], [62], [60], [79], [83], [135], lo que hemos dado en llamar distorsión euclídea para diferenciarla de otras funciones de distorsión más generales:

$$\xi_E^2 = \sum_{i=1}^{c} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mathbf{y}_i\|^2 , \tag{2.61}$$

donde c es el número de clases, $\mathbf{y}_i$ es el centro de la i-esima clase. Como ya hemos mencionado en la sección 2.2 la minimización de la distorsión euclídea corresponde

a la estimación máximo verosimil de los parámetros de la distribución mezcla asumiendo que las d.d.p. componente siguen una distribución normal con matriz de covarianza identidad. Además, S_i es la partición de la muestra correspondiente a la i-esima clase. Esta función objetivo es la misma que la minimizada por el algoritmo de K-medias.

La distorsión euclídea puede escribirse como

$$\xi_E^2(\mathbf{Y},\mathbf{x}) = \sum_{i=1}^{c} E_{\mathbf{x}}\left[\|\mathbf{x}-\mathbf{y}_i\|^2 \,|\mathbf{x}\in\mathcal{R}_i\right] \tag{2.62}$$

donde $\mathcal{R}_i$ denota la región asociada a la clase ω_i representada por $\mathbf{y}_i$. El vector $\mathbf{Y}$ denota el conjunto de todos los representantes (medias) de las clases. $E_{\mathbf{x}}[.]$ denota la esperanza sobre el dominio de definición y la d.d.p. de la variable aleatoria $\mathbf{x}$. Dado que los centros de clases son independientes funcionalmente obtenemos

$$\frac{\partial \xi_E^2(\mathbf{y}_i,\mathbf{x})}{\partial \mathbf{y}_i} = \sum_{i=1}^{c} E_{\mathbf{x}}\left[-2(\mathbf{x}-\mathbf{y}_i)\,|\mathbf{x}\in\mathcal{R}_i\right] \quad 1\le i\le c \tag{2.63}$$

y el gradiente instantáneo tiene la forma

$$\left.\frac{\partial}{\partial \mathbf{y}_i}\xi_E^2\right|_{\mathbf{x}} = -2(\mathbf{x}-\mathbf{y}_i)\,\delta_i(\mathbf{x},\mathbf{Y}). \tag{2.64}$$

Aplicando el método de Robins Monro para minimizar la distorsión euclídea

$$\mathbf{y}_i(t+1) = \mathbf{y}_i(t) - a(t)\left.\frac{\partial}{\partial \mathbf{y}_i}\xi_E^2\right|_{\mathbf{x}(t)}, \tag{2.65}$$

obtenemos el algoritmo de Aprendizaje Competitivo Simple (SCL)

$$\mathbf{y}_i(t+1) = \mathbf{y}_i(t) + a_t(\mathbf{x}(t) - \mathbf{y}_i(t))\,\delta_i(\mathbf{x}(t),\mathbf{Y}) \tag{2.66}$$

donde los coeficientes de pertenencia especifican la función discriminante que clasifica los vectores de características

$$\delta_i(\mathbf{x},\mathbf{Y}) = \begin{cases} 1 & i = \underset{k=1..,c}{argmin}\left\{\|\mathbf{x}-\mathbf{y}_k\|^2\right\} \\ 0 & sino \end{cases}, \tag{2.67}$$

y que, como ya se ha dicho, es equivalente a las funciones discriminantes dadas por los logaritmos de las probabilidades *a posteriori*.

La convergencia del SCL al libro de códigos óptimo dependerá de las condiciones iniciales, dada la naturaleza local del algoritmo de descenso de gradiente. La calidad de los resultados también dependerá de la velocidad de aprendizaje $a(t)$ cuya programación se discute más adelante en términos generales.

2.6.2.La regla competitiva general

Pueden considerarse en general como métodos de agrupamiento basados en el descenso por el gradiente estocástico que tratan de minimizar funciones objetivo particulares. La más inmediata es la dispersión o distorsión Euclídea que da lugar a la regla Competitiva Simple y que es el equivalente estocástico al algoritmo Isodata o algoritmo de las K-medias. Revisaremos algunas de las más conocidas, pero nuestro interés está en el algoritmo de los mapas auto-organizativos (SOM) que es la regla que vamos a aplicar como método de estimación de los libros de códigos (codebooks) que utilizaremos en nuestras aplicaciones prácticas.

El SCL es el caso más sencillo de la expresión general de la regla de aprendizaje para RNC la cual tiene la siguiente forma general (t es el orden de presentación de los vectores input).

$$\triangle \mathbf{y}_k(t) = a(t) \cdot V_k(\mathbf{Y}, \mathbf{x}_t, t) \cdot (\mathbf{x}_t - \mathbf{y}_k(t)) \quad k = 1, .., c \tag{2.68}$$

donde $a(t)$ es la velocidad de aprendizaje o factor de ganancia y $V_k(\mathbf{Y}, \mathbf{x}_t, t)$ es la función de vecindad que define el conjunto de unidades que se adaptan junto con la vencedora al presentarse un vector intput. Esta función de vecindad varía de un tipo de red a otra y condiciona fuertemente las propiedades de la red relativas a su convergencia durante el aprendizaje, estabilidad numérica y propiedades de la red una vez entrenada. La velocidad de aprendizaje es un factor numérico que obedece a las condiciones que se imponen a los algoritmos de tipo Robins-Munro para garantizar su convergencia.

Si consideramos que la función de vecindad permanece fija durante el proceso de adaptación

$$V_k(\mathbf{Y}, \mathbf{x}_t, t) = V_k(\mathbf{Y}, \mathbf{x}); \forall t. \tag{2.69}$$

Podríamos hipotetizar que la regla de aprendizaje es un proceso de descenso del gradiente estocástico, que realiza la minimización de alguna función objetivo. Esto significa que estamos asumiendo que la función de vecindad es proporcional al opuesto del gradiente instantaneo de la hipotética función objetivo

$$V_k(\mathbf{Y}, \mathbf{x}) \propto -\left. \frac{\partial}{\partial \mathbf{y}_k} \xi_V \right|_{\mathbf{x}}, \tag{2.70}$$

donde el factor de proporcionalidad es el vector diferencia $(\mathbf{x}_t - \mathbf{y}_k(t))$ y que, bajo las condiciones apropiadas, esta función puede ser deducida de la función de

vecindad como sigue:

$$\begin{aligned} \xi_V &= E_{\mathbf{x}}\left[\sum_{k=1}^{c}\int -V_k\left(\mathbf{Y},\mathbf{x}\right)\left(\mathbf{x}-\mathbf{y}_k\right)d\mathbf{y}_k\right] \\ &= \sum_{k=1}^{c}\int\int -V_k\left(\mathbf{Y},\mathbf{x}\right)\left(\mathbf{x}-\mathbf{y}_k\right)p\left(\mathbf{x}\right)d\mathbf{y}_k d\mathbf{x}. \end{aligned} \tag{2.71}$$

Mientras $V_k\left(\mathbf{Y},\mathbf{x}\right) \neq \delta_k\left(\mathbf{x},\mathbf{Y}\right)$ esta función objetivo (ecuación 2.71) será diferente de la distorsión Euclídea. El interés de una función objetivo particular dependerá de las características del agrupamiento buscadas para una aplicación particular.

Usualmente las funciones de vecindad varían durante el proceso de adaptación o aprendizaje. Esto es, son dependientes del tiempo de adaptación. La forma analítica de la función minimizada viene a ser entonces mucho más compleja, en muchos casos permanece desconocida. Una aproximación conveniente es ver la regla de aprendizaje (2.68) como realizando una cascada de minimizaciones sobre una secuencia de funciones objetivo.

$$\xi_V\left(t\right) = \sum_{k=1}^{c}\int\int -V_k\left(\mathbf{Y},\mathbf{x},t\right)\left(\mathbf{x}-\mathbf{y}_k\right)p\left(\mathbf{x}\right)d\mathbf{y}_k d\mathbf{x}. \tag{2.72}$$

El límite de esta secuencia de funciones objetivo será determinado por el límite de sus respectivas funciones de vecindad:

$$\lim_{t\to\infty} V_k\left(\mathbf{Y},\mathbf{x},t\right) = V_k^*\left(\mathbf{Y},\mathbf{x}\right) \Longrightarrow \lim_{t\to\infty} \xi_V\left(t\right) = \xi_{V^*}. \tag{2.73}$$

La aplicación de la ecuación 2.68 es, por tanto, un procedimiento encaminado a minimizar la función objetivo límite. Este procedimiento se supone que añade alguna propiedad interesante al puro descenso de gradiente de ξ_{V^*}. La propiedad más deseada es la de minimización global. Se espera que la aplicación de la regla competitiva general (2.68) producirá una minimización global de la función objetivo límite ξ_{V^*}. Este proceso recuerda al del enfriamiento estadístico (*simulated annealing*) [1]. Los parámetros de control son los parámetros que determinan la anchura y forma de la función de vecindad.

En la aplicación más común de la regla dada en la ecuación 2.68 la función de vecindad se encoje hasta convertirse en la función de pertenencia dura:

$$\lim_{t\to\infty} V_k\left(\mathbf{Y},\mathbf{x},t\right) = \delta_k\left(\mathbf{x},\mathbf{Y}\right). \tag{2.74}$$

Por tanto, se puede asumir que la secuencia de funciones objetivo converge a la distorsión Euclidea

$$\lim_{t\to\infty} \xi_V(t) \approx \xi_E^2, \tag{2.75}$$

o a alguna otra función equivalente que tiene sus mínimos en los mismos lugares. Esta convergencia se controla por parámetros específicos de la función de vecindad. En lugar de estudiar la convergencia funcional de la función objetivo, podemos estudiar los casos límite de la función de pertenencia. Discutiremos para el SOM, FLVQ y SCS su convergencia al SCL que minimiza la distorsión euclídea. Bajo esta visión, SOM, FLVQ y SCS son inicializaciones robustas para el SCL. En otras palabras, deben funcionar como minimizadores globales de la distorsión euclídea.

Esta interpretación contrasta con las interpretaciones de las funciones de vecindad como modificaciones de la velocidad de aprendizaje $a_i(t)$, como se discute en [8], [153]. Consideramos que la velocidad de aprendizaje debe tener la misma secuencia de valores o programación en todos los casos, y que los restantes coeficientes caracterizan la secuencia de funciones objetivo que están siendo minimizadas.

2.6.3.Mapa Auto-Organizativo (*Self-Organizing Map*)

La idea fundamental de los mapas de características auto-organizados fue introducida originalmente por Malsburg [94] y Grossberg [56] para explicar la formación de mapas topológicos neuronales. Basado en estos trabajos, Kohonen [78], [79], [80] propone su modelo de RNC denominado Mapa Auto-Organizativo (SOM) como un mecanismo de aprendizaje no supervisado inspirado en la correspondencia topológica entre la localización en el cortex cerebral de las sensaciones tactiles y la distribución espacial en el cuerpo del origen sensorial. Esta correspondencia topológica se manifiesta en la proximidad en el cortex de los receptores correspondientes a partes del cuerpo cercanas. La representación más famosa es homúnculo que se extiende sobre la corteza cerebral. Esta correspondencia se extiende de otras maneras a otras formas sensorias, en las que estimulos similares se localizan en lugares cercanos del cortex cerebral. SOM se ha aplicado con éxito a un gran número de problemas ingenieriles de todo ámbito. Otros algoritmos se han propuesto en la literatura basados en la idea original, entre ellos las estructuras de crecimiento de Fritzke [37] [38].

Se considera un espacio de características $\mathbb{R}^N$ con vectores de características $\mathbf{x} \in \mathbb{R}^N$ y representantes de las clases $\mathbf{y}_i \in \mathbb{R}^N$, y un espacio de los índices $I \subset \mathbb{N}^d$ sobre el que también esta definida una distancia $|i-k| \geq 0; i,k \in I$. Existe por

tanto una organización propia de las neuronas en su disposición espacial independiente de su "contenido", esto es, de los pesos. Por otro lado, el espacio de características tiene a su vez una topología y una métrica asociadas, usualmente la distancia Euclídea.

La propiedad de conservación topológica se puede expresar mediante las siguientes relaciones, válidas para todo $i, k, l \in I$:

$$\|\mathbf{y}_i - \mathbf{y}_k\|^2 = \min_{l \neq i} \left\{ \|\mathbf{y}_i - \mathbf{y}_l\|^2 \right\} \Rightarrow |i-k| = \min_{l \neq i} \{|i-l|\}, \tag{2.76}$$

$$\|\mathbf{y}_i - \mathbf{y}_k\|^2 \geq \|\mathbf{y}_i - \mathbf{y}_l\|^2 \Rightarrow |i-k| \geq |i-l|. \tag{2.77}$$

La implicación 2.76 indica que si los pesos de las unidades están próximos en el espacio de características las correspodientes unidades deben ser también vecinas en el espacio de los índices. La implicación 2.77 especifica que el orden relativo de los vectores input debe preservarse en el orden relativo de los correspondientes índices de las unidades. La preservación topológica tiene implicaciones en las propiedades numéricas de los mapas auto-organizativos, una vez que han sido "correctamente" entrenados.

Si consideramos el caso $d = N$ en el que el espacio input y el de los indices tienen la misma dimension, se puede considerar que los pesos de las unidades del SOM están realizando una discretización del espacio de características guiada por la distribución de probabilidad de los inputs presentados a la red durante su entrenamiento.

Si $d < N$ y se ha conseguido entrenar una red que cumple las condiciones de preservación topológica, obtenemos una proyección no lineal del espacio input en el espacio de los indices. Esta proyección constituye una técnica de reducción de dimensiones muy poderosa.

La regla de aprendizaje sigue la forma de las reglas competitivas

$$\triangle \mathbf{y}_k(t) = a(t) \cdot V_k(\mathbf{Y}, \mathbf{x}_t, t) \cdot (\mathbf{x}_t - \mathbf{y}_k(t)) \quad k = 1, .., c \tag{2.78}$$

donde $V_k(\mathbf{Y}, \mathbf{x}_t, t)$ es la función de vecindad que, en el caso del SOM, está definida sobre la topología de los índices de las unidades

$$V_k(\mathbf{Y}, \mathbf{x}_t, t) = v_t(k, i) \text{ donde } i = \arg\min_{k=1,..,c} \left\{ \|\mathbf{x} - \mathbf{y}_k\|^2 \right\} \tag{2.79}$$

La notación anterior es lo sufientemente general como para abarcar todos los casos de reglas competitivas. En el caso del SCL tenemos que $v_t(k, i) = \delta_k(i)$. En

el caso del Soft-Competition (SCS) $V_k(\mathbf{Y}, \mathbf{x}_t, t) = \widehat{p}(\omega_k | \mathbf{x}_t)$. El SOM se distingue por que la función de vecindad depende de la topología de los índices. Algunas posibles funciones de vecindad son las siguientes:

1. Vecindad discreta
$$v_t(k,i) = \begin{cases} 1 & |k-i| \leq \varepsilon(t) \\ 0 & |k-i| > \varepsilon(t) \end{cases} \tag{2.80}$$

2. Vecindad continua
$$v_t(k,i) = f_t(|k-i|) \tag{2.81}$$
donde f_t es cualquier función, las mas comunes son gausianas o derivadas de gausianas
$$f_t(x) = e^{-x^2/2\sigma^2(t)} \tag{2.82}$$
donde el parámetro de velocidad de aprendizaje es decreciente con el tiempo, sin llegar a anularse. Una de las mas empleadas [100] es
$$C(i,j,k) = \exp\left(\frac{-\|i-j\|^2}{(\beta(k) S_0)^2}\right), \tag{2.83}$$
donde $\beta(k)$ es la velocidad de aprendizaje y S_0 el Número de unidades en cada dimension del mapa. En ocasiones la velocidad de aprendizaje se hace particular para cada neurona o vector código. En este caso la regla de adaptación toma la forma
$$\triangle \mathbf{y}_k(t) = \alpha_{k,t} \cdot V_k(\mathbf{Y}, \mathbf{x}_t, t) \cdot (\mathbf{x}_t - \mathbf{y}_k(t)) \quad k = 1, .., c \tag{2.84}$$
donde $\alpha_{k,t}$ es la velocidad de aprendizaje local para cada vector código, que decrece a cero en la forma habitual para el algoritmo de gradiente estocástico de forma independiente para cada vector código. Una forma de definir esta velocidad de aprendizaje local puede ser
$$\alpha_{k,t} = \alpha_0(1 - \tau_k/n) \tag{2.85}$$
con
$$\tau_k = \sum_{i=1}^{t} V_k(\mathbf{Y}, \mathbf{x}_i, i), \tag{2.86}$$
de forma que la velocidad de aprendizaje será sensible al tamaño del agrupamiento asociado con cada vector código. Esto es, τ_k acumula el número de veces que el vector código $\mathbf{y}_k$ ha sido actualizado ya sea por qué es el

vencedor o por qué la función de vecindad fuerza su adaptación, si la función de vecindad es de la forma

$$V_k(\mathbf{Y}, \mathbf{x}_t, t) = \begin{cases} 1 & |i - j(\mathbf{x}_t)| \leq v_t \\ 0 & \text{otherwise} \end{cases} \tag{2.87}$$

donde $j(\mathbf{x}_t)$ denota el índice del vector código más cercano a la muestra $\mathbf{x}_t$.

$$j(\mathbf{x}_t) = \underset{k=1,..,c}{\arg\min} \left\{ \|\mathbf{x} - \mathbf{y}_k\|^2 \right\} \tag{2.88}$$

Notese que, en los límites,

$$\begin{aligned} \lim_{v \to \infty} V_k(\mathbf{Y}, \mathbf{x}_t, t) &= 1 \\ \lim_{v \to 0} V_k(\mathbf{Y}, \mathbf{x}_t, t) &= \delta_i(\mathbf{x}, \mathbf{Y}) \end{aligned} \tag{2.89}$$

la función de vecindad pasa de ser perfectamente borrosa cuando el radio de vecindad es infinito a ser la función de pertenencia dura o crisp cuando es cero, lo que corresponde a la convergencia al SCL.

Si consideramos que la función de vecindad es invariante en el tiempo $v_t(k,i) = v(k,i)$ se puede deducir [62], [75], [79], [23] la siguiente función de energía

$$\xi_{SOM} = \sum_{i=1}^{c} \int \int -V_i(\mathbf{Y}, \mathbf{x}_t, t)(\mathbf{x} - \mathbf{y}_i) p(\mathbf{x}) d\mathbf{y}_i d\mathbf{x} \tag{2.90}$$

$$= \sum_{i=1}^{c} \int_{\bigcup_{j \in V(i,v)} R_j} \|\mathbf{x} - \mathbf{y}_i\|^2 p(\mathbf{x}) d\mathbf{x} \tag{2.91}$$

$$= \sum_{i=1}^{c} E_{\mathbf{x}} \left[\|\mathbf{x}_j - \mathbf{y}_i\|^2 \middle| \mathbf{x} \in \bigcup_{j \in V(i,v)} R_j \right] \tag{2.92}$$

o si consideramos que la función de vecindad es general (no una función escalón unitario) tenemos

$$\xi_{SOM} = \sum_k E_{\mathbf{x}} \left[v(k,i) \|\mathbf{x} - \mathbf{y}_k\|^2 | \mathbf{x} \in \mathcal{X}_k \right] \tag{2.93}$$

que viene a ser una extensión de la distorsión euclídea en la que cada vector código contribuye a la distorsión debida a la cuantización de su teselación del espació input y a la de sus codevectores vecinos.

En general esta función de energía nos dice muy poco a cerca de la convergencia del SOM y el caso de vecindad constante no tiene interes practico.

El radio de vecindad disminuye conforme avanza el tiempo de la adaptación:

$$V_i(\mathbf{x},\mathbf{Y},t) = \begin{cases} 1 & |j(\mathbf{x}_t) - i| \leq v(t) \\ 0 & otherwise \end{cases} ; 1 \leq i \leq c \tag{2.94}$$

El vecindario decrece gradualmente, por lo que

$$\lim_{\tau\to\infty} v(t) = 0 \Longrightarrow \lim_{\tau\to\infty} V_i(\mathbf{x},\mathbf{Y},t) = \delta_i(\mathbf{x},\mathbf{Y}) \tag{2.95}$$

y por tanto podríamos asumir que existe una convergencia funcional de la función minimizada por el SOM hacia la distorsión euclídea:

$$\lim_{\tau\to\infty} \xi_{SOM}(t) \approx \xi_E^2 \tag{2.96}$$

La convergencia funcional depende de los parámetros que controlan la extensión de la función de vecindad. El análisis formal de la convergencia del SOM está centrado en la convergencia a estados organizados [22], [32], [33], [34]. Algunas reglas de aprendizaje inspiradas en el SOM [139], [140], [141], [133] se proponene con el único objetivo de alcanzar propiedades de convergencia deseables hacia estados organizados. En nuestra perspectiva, la convergencia a estados organizados es sólo significativa si dichos estados aseguran una solución mejor al problema de la minimización de la distorsión euclídea, esto es si mejoran al SCL en este sentido tanto en robustez frente a las condiciones iniciales como en soluciones puntuales.

2.6.4.Esquema de Competición Suave (SCS)

La primera referencia que hemos encontrado en la literatura al método de Competición Suave (Soft Competition) (SCS) es [153], donde el SCS se propone como una combinación del enfriamiento simulado (*simulated annealing* [1]) y el SCL. Sin embargo, la definición de los parámetros de aprendizaje es muy obscura y su programación díficil de ajustar. En [8] se revisita el algoritmo dando una interpretación estadística de los coeficientes de vecindad. Se muestra que corresponden a las probabilidades a posteriori de las clases dado el input, asumiendo que el modelo de la distribución de los datos es una mezcla de gausianas. Esta interpretación se refuerza por la derivación del SCS a partir de la entropía cruzada de Kullback-Leibler que se detalla a continuación.

La regla de aprendizaje del SCS se deduce cuando se pretende minimizar mediante un algoritmo de descenso de gradiente la distancia de Kullback-Leiber entre la distribución empírica de los datos $p(\mathbf{x})$, calculada a partir de una muestra $\aleph = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, y el modelo $p'(\mathbf{x})$ dado por una densidad mezcla de densidades normales

$$D(p, p') = \int_{\mathbf{x}} p(\mathbf{x}) \, ln \frac{p(\mathbf{x})}{p'(\mathbf{x})} d\mathbf{x}, \tag{2.97}$$

donde las densidades componente del modelo son $N(\mathbf{y}_i, \sigma_i^2 \mathbf{I})$

$$p'(\mathbf{x}) = \frac{1}{c} \sum_{i=1}^{c} \frac{1}{(2\pi)^{d/2} \sigma_i^d} e^{-\frac{\|\mathbf{x}-\mathbf{y}_i\|^2}{2\sigma_i^2}}. \tag{2.98}$$

Si queremos ser consistentes en la notación, que la distancia de Kullback-Leibler ocupa el lugar de la distorsión que pretendemos minimizar:

$$\xi_{SCS} \equiv D(p, p'). \tag{2.99}$$

Aplicando el método del gradiente estocástico a las medias y a las varianzas de las clases

$$\triangle \mathbf{y}_i(t) = a_t \frac{\partial}{\partial \mathbf{y}_i} D(p, p') \Big|_{\mathbf{x}(t)} \qquad i = 1, .., c. \tag{2.100}$$

$$\triangle \sigma_i(t) = b_t \frac{\partial}{\partial \sigma_i} D(p, p') \Big|_{\mathbf{x}(t)} \qquad i = 1, .., c. \tag{2.101}$$

Si consideramos el caso en el que las varianzas de todas las densidades gausianas son idénticas $\sigma_i^2 = \sigma^2$ para $i = 1, .., c$, el gradiente instantaneo de la entropía cruzada 2.97, relativo a las medias de las densidades gausianas es

$$\frac{\partial}{\partial \mathbf{y}_i} D(p, p') \Big|_{\mathbf{x}} = -\frac{1}{\sigma^2} \frac{e^{-\frac{\|\mathbf{x}-\mathbf{y}_i\|^2}{2\sigma^2}}}{\sum_{k=1}^{c} e^{-\frac{\|\mathbf{x}-\mathbf{y}_k\|^2}{2\sigma^2}}} (\mathbf{x} - \mathbf{y}_i) \tag{2.102}$$

$$= -\frac{1}{\sigma^2} P_i(\mathbf{x}(t)) (\mathbf{x} - \mathbf{y}_i) \tag{2.103}$$

Un gradiente similar se puede derivar para la varianza de las distribuciones gausianas:

$$\frac{\partial}{\partial \sigma} D(p, p') \Big|_{\mathbf{x}} = -P_i(\mathbf{x}(t)) \frac{1}{\sigma^3} \left(\sum_{i=1}^{c} \|\mathbf{x} - \mathbf{y}_i\|^2 - d\sigma^2 \right) \tag{2.104}$$

Esto es, en el caso de varianzas idénticas para todas las clases obtenemos las reglas competitivas

$$\triangle \mathbf{y}_i(t) = \frac{a_t}{\sigma^2} P_i(\mathbf{x}(t))(\mathbf{x}(t) - \mathbf{y}_i(t)), \tag{2.105}$$

$$\triangle \sigma(t) = \frac{b_t}{\sigma^3} \sum_{i=1}^{c} P_i(\mathbf{x}(t)) \left(\|\mathbf{x}(t) - \mathbf{y}_i(t)\|^2 - d\sigma(t) \right), \tag{2.106}$$

donde $d = dim(\mathbf{x})$ y los coeficientes de pertenencia son aproximaciones a las probabilidades *a posteriori* de la forma que funcionan como vecindades

$$V_i(\mathbf{Y}, \mathbf{x}_t, t, \sigma) = P_i(\mathbf{x}) = \frac{\frac{1}{N} e^{-\frac{\|\mathbf{x}-\mathbf{y}_i\|^2}{2\sigma^2}}}{\sum_{j=1}^{c} e^{-\frac{\|\mathbf{x}-\mathbf{y}_j\|^2}{2\sigma^2}}}. \tag{2.107}$$

Como en otros casos de reglas competitivas que estamos tratando, es posible demostrar que

$$\begin{aligned} \lim_{\sigma \to \infty} V_i(\mathbf{Y}, \mathbf{x}_t, t, \sigma) &= \frac{1}{c}, \\ \lim_{\sigma \to 0} V_i(\mathbf{Y}, \mathbf{x}_t, t, \sigma) &= \delta_i(\mathbf{x}, \mathbf{Y}), \end{aligned} \tag{2.108}$$

y que, por tanto, el SCS converge a la regla SCL para valores especificos del parámetro que controla la extensión del vecindario.

En las definiciones previas, hemos asumido el modelo más sencillo. Este modelo se puede hacer más complejo permitiendo que cada componente de la densidad mezcla 2.98 tenga distintas varianzas isotrópicas σ_i^2, diferentes varianzas en cada eje de representación de los datos σ_{ij}^2, e incluso matrices de covarianza no diagonales Σ_i. Estas distintas asunciones dan lugar a variaciones en la regla de adaptación de la varizanza (2.106) pero afectan trivialmente a la regla de adaptación de los vectores código (2.105)

Hemos aplicado las reglas para los vectores código (2.105) y para sus varianzas asociadas (2.106) a la tarea de Agrupamiento no estacionario (*Non Stationary Clustering*) en [48]. Sin embargo, los resultados no son comparables con los de otros algoritmos competitivos probados debido a que estas dos reglas cuando se aplican simultaneamente corresponden a la minimización de una secuencia de funciones que no converge a la distorsión euclídea. Puesto que la regla en la

ecuación 2.106 hace que la varianza asociada a los vectores código converja a un valor no nulo

$$\lim_{t\to\infty} \sigma(t) = \sigma^* \tag{2.109}$$

la función de vecindad converge a una función distinta de la función de pertenencia dura

$$\lim_{t\to\infty} V_k(\mathbf{Y}, \mathbf{x}_t, t, \sigma_t) = V_k(\mathbf{Y}, \mathbf{x}_t, t, \sigma^*) \neq \delta_i(\mathbf{x}, \mathbf{Y}), \tag{2.110}$$

lo que implica que la función objetivo que se está minimizando en el límite no coincide con la distorsión euclídea

$$\lim_{t\to\infty} D(\aleph, \mathbf{Y}, \sigma) = D(\aleph, \mathbf{Y}, \sigma^*) \neq \widehat{\xi}_E^2. \tag{2.111}$$

Hemos de tener en cuenta que otros algoritmos reducen su función de vecindad hasta que se convierte en la función de pertenencia dura mediante la programación de sus parámetros de control. Es importante también notar que la medida de comparación de rendimiento natural entre algoritmos, salvo que se consideren propiedades como el orden topológico de las unidades neuronales, es la distorsión euclídea.

Si ponemos las cosas en el marco de la cuantización vectorial, la estimación de la varianza optima en el sentido de minimizar (2.97) introduce la distancia de Mahalanobis [28], [135] como la distancia de error, por lo que el cuantizador del vecino más cercano debería ser definido de acuerdo a la distancia de Mahalanobis. La arquitecura HEC [95] produce como resultado del entrenamiento un VQ basado en la distancia (regularizada) de Mahalanobis . Este podría ser también el caso con SCS si las funciones de vecindad se usan para realizar la cuantización.

Por todo ello, en nuestros trabajos, la varianza de las densidades gausianas se decrece para obtener la convergencia funcional deseada. Esto es, para hacer que el SCS minimize realmente la distorsión euclídea. Por tanto, definimos el vecindario dependiente del tiempo como:

$$V_i(\mathbf{Y}, \mathbf{x}_t, t) = \frac{e^{-\frac{\|\mathbf{x}-\mathbf{y}_i\|^2}{2\sigma(t)^2}}}{\sum_{k=1}^{c} e^{-\frac{\|\mathbf{x}-\mathbf{y}_k\|^2}{2\sigma(t)^2}}}; 1 \leq i \leq c \tag{2.112}$$

y el proceso comienza desde grandes valores de la varianza asumida en el modelo y se decrece hasta cero, de forma que

$$\lim_{t\to\infty} \sigma(t) = 0 \Longrightarrow \lim_{t\to\infty} V_i(\mathbf{Y}, \mathbf{x}_t, t) = \delta_i(\mathbf{x}, \mathbf{Y}), \tag{2.113}$$

$$\Longrightarrow \lim_{t\to\infty} \xi_{SCS}(t) \approx \xi_E^2. \tag{2.114}$$

Este proceso es muy sensible a la programación de las varianzas. Adaptaciones muy largas con valores altos de la varianzas colapsa los vectores código en la media de la muestra, una situación irrecuperable. Hemos observado que, si la convergencia a SCL es muy rápida, no se producirán mejoras respecto de SCL.

2.6.5.Neural Gas

La red Neural Gas se caracteriza por la función siguiente vecindad

$$V_k\left(\mathbf{Y}, \mathbf{x}_t, t\right) = e^{\left(-\frac{1}{\lambda(t)} rank_i(x,Y)\right)} \tag{2.115}$$

donde $rank_i\left(\mathbf{x}, \mathbf{Y}\right)$ es el orden del codevector $\mathbf{y}_i$ al presentarse $\mathbf{x}$. No hemos realizado trabajos con esta red en el marco de esta tesis, si bien en [96] se demostraba que proporcionaba mejores resultados que el SOM en algunos problemas benchmark.

2.6.6.Cuantización Vectorial Borrosa (FLVQ)

La Cuantización Vectorial Borrosa (Fuzzy Learning Vector Quantization) (FLVQ) se deriva como la minimización del criterio borroso de agrupamiento (*Fuzzy Clustering*) (2.56), originalmente propuesto para justificar la generalización borrosa del algoritmo de las K-medias formulando el algoritmo borroso de las c-medias (Fuzzy c-means). El proceso del FLVQ es una secuencia de minimizaciones a diferentes valores del exponente m. En este caso las expresión final del criterio borroso de agrupamiento es la distorsión euclídea, y FLVQ se puede interpretar como un minimizador de la distorsión euclídea. Existe una gran cantidad de literatura sobre algoritmos relacionados de alguna manera con FLVQ [8], [74], [72], [136], y bastante debate sobre su apropiada aplicación como un algoritmo online [18], [108], [111]. Nosotros lo consideramos como un algoritmo online dentro de la clase caracterizada por la regla competitiva general (2.68).

Recordamos la función que expresa el criterio borroso de agrupamiento:

$$J_m\left(U, \mathbf{Y}\right) = \sum_{j=1}^{n} \sum_{i=1}^{c} \left(u_{ij}\right)^m \left\|\mathbf{x}_j - \mathbf{y}_i\right\|_A^2 \tag{2.116}$$

donde $\|.\|_A^2$ denota una distancia euclídea pesada por una matriz A y los coefi-

cientes borrosos de pertenencia son similares a los definidos en (2.57):

$$u_{ij} = \left(\sum_{k=1}^{c} \left(\frac{\|\mathbf{x}_j - \mathbf{y}_i\|_A^2}{\|\mathbf{x}_j - \mathbf{y}_k\|_A^2} \right)^{\frac{1}{m-1}} \right)^{-1} . \tag{2.117}$$

Estos coeficientes especifican la partición borrosa de la muestra.

La regla competitiva en este caso tiene como función de vecindad precisamente estos coeficientes borrosos de pertenencia

$$V_i(\mathbf{Y}, \mathbf{x}_t, t, m) = \left(\sum_{k=1}^{c} \left(\frac{\|\mathbf{x} - \mathbf{y}_i\|^2}{\|\mathbf{x} - \mathbf{y}_k\|^2} \right)^{\frac{1}{m(t)-1}} \right)^{-m(t)} ; 1 \leq i \leq c \tag{2.118}$$

donde el parámetro $m(t)$ controla el grado de borrosidad y tiende a 1 por arriba. Esta función de vecindad está bien definida cuando el vector input no coincide con ningún vector código:

$$\mathbf{x} \neq \mathbf{y}_i; \forall i. \tag{2.119}$$

La solución práctica a esta singularidad es ignorar aquellos vectores input que coinciden con algún vector código. Se puede mostrar [8] que esta función de vecindad se convierte en la función de pertenencia dura cuando m se acerca a 1 por arriba.

$$\lim_{m \to 1^+} u_i(\mathbf{x}, \mathbf{Y}) = \delta_i(\mathbf{x}, \mathbf{Y}) \Longrightarrow \lim_{m \to 1^+} V_i(\mathbf{Y}, \mathbf{x}_t, t, m) = \delta_i(\mathbf{x}, \mathbf{Y}), \tag{2.120}$$

de forma que la regla de adaptación de la ecuación 2.68, con la función de vecindad dada por la ecuación 2.118 se convierte en SCL cuando m se acerca a 1 por arriba. Cuando m crece, la función de vecindad se hace más borrosa, todos los vectores código son afectados de la misma manera por el input, pero está amortigüada exponencialmente

$$\lim_{m \to \infty} u_i(\mathbf{x}, \mathbf{Y}) = \frac{1}{c} \Longrightarrow \lim_{m \to \infty} V_i(\mathbf{Y}, \mathbf{x}_t, t, m) = 0, \tag{2.121}$$

Valores muy altos de m no producen adaptaciónes.

La función minimizada manteniendo m constante está relacionada con el criterio borroso de agrupamiento, aunque FLVQ no se puede derivar como un algoritmo de descenso de gradiente estocástico del criterio de agrupamiento borroso dado en la ecuación 2.56. De hecho, la expresión analítica de la función minimizada por

la aplicación de las reglas 2.68 y 2.118, para valor fijo de m, sería el resultado de los siguientes cálculos no triviales:

$$\begin{aligned}\xi_{FLVQ} &= \sum_{i=1}^{c} \int \int -V_i\left(\mathbf{Y}, \mathbf{x}_t, t, m\right)\left(\mathbf{x}-\mathbf{y}_i\right) p\left(\mathbf{x}\right) d\mathbf{y}_i d\mathbf{x} \qquad (2.122)\\ &= \sum_{i=1}^{c} E_{\mathbf{x}}\left[\int -\left(\sum_{k=1}^{c}\left(\frac{\left\|\mathbf{x}-\mathbf{y}_i\right\|^2}{\left\|\mathbf{x}-\mathbf{y}_k\right\|^2}\right)^{\frac{1}{m-1}}\right)^{-m}\left(\mathbf{x}-\mathbf{y}_i\right) d\mathbf{y}_i\right],\end{aligned}$$

donde $E_{\mathbf{x}}\left[.\right]$ denota la expectación en $\mathbf{x}$.

No estamos interesados en conocer la expresión exacta de la función objetivo minimizada, en tanto que seamos capaces de deducir su convergencia funcional cualitativa a partir del examen de la función de vecindad. Hemos encontrado sólo un intento de derivar la expresión exacta del gradiente de la función criterio de agrupamiento borroso de la ecuación 2.56 en [117]. Hemos encontrado que esta derivación es incorrecta en alguno de sus pasos, por ello hemos derivado la expresión correcta en el Apéndice D. Esta expresión es la siguiente:

$$\frac{\partial}{\partial \mathbf{y}_i} J_m\left(U, \mathbf{Y}, \aleph\right) = -\frac{2m}{m-1}\sum_{j=1}^{n}\left\{\sum_{k=1}^{c}\left[\left(u_{kj}\right)^{m+1}\left(\frac{\left\|\mathbf{x}_j-\mathbf{y}_k\right\|^2}{\left\|\mathbf{x}_j-\mathbf{y}_i\right\|^2}\right)^{\frac{m}{m-1}}\right]-\frac{u_{ij}^m}{m}\right\}\left(\mathbf{x}_j-\mathbf{y}_i\right). \qquad (2.123)$$

Esta expresión conduce a una regla competitiva con la siguiente función de vecindad

$$\Phi_i^{(J_m)}\left(\mathbf{x}, \mathbf{Y}, m\right) = \frac{m}{m-1}\left\{\sum_{k=1}^{c}\left[\left(u_k\left(\mathbf{x}, \mathbf{Y}\right)\right)^{m+1}\left(\frac{\left\|\mathbf{x}-\mathbf{y}_k\right\|^2}{\left\|\mathbf{x}-\mathbf{y}_i\right\|^2}\right)^{\frac{m}{m-1}}\right]-\frac{\left(u_i\left(\mathbf{x}, \mathbf{Y}\right)\right)^m}{m}\right\}, \qquad (2.124)$$

que muestra los mismos valores límite que la función de vecindad dada en la ecuación 2.118:

$$\begin{aligned}\lim_{m\to 1^+} \Phi_i^{(J_m)}\left(\mathbf{x}, \mathbf{Y}, m\right) &= \delta_i\left(\mathbf{x}, \mathbf{Y}\right), \qquad (2.125)\\ \lim_{m\to\infty} \Phi_i^{(J_m)}\left(\mathbf{x}, \mathbf{Y}, m\right) &= \lim_{m\to\infty}\left(\frac{1}{c}\right)^m = 0,\end{aligned}$$

pero que es mucho más compleja en su cáculo. Por tanto asumimos que la función de vecindad 2.118 es un representante conveniente de la familia completa de aproximaciones on-line al agrupamiento borroso.

La función de vecindad dependiente del tiempo se define mediante la programación en el tiempo del valor del parámetro de control

$$V_i\left(\mathbf{Y},\mathbf{x}_t,t\right)=\left(\sum_{k=1}^{c}\left(\frac{\left\|\mathbf{x}-\mathbf{y}_i\right\|^2}{\left\|\mathbf{x}-\mathbf{y}_k\right\|^2}\right)^{\frac{1}{m(t)-1}}\right)^{-m(\tau)} ;1\leq i\leq c. \tag{2.126}$$

La programación comienza con valores altos de $m\left(0\right)$ para decrecer hasta 1.

$$\lim_{t\to\infty} m\left(t\right) = 1^{+}\Longrightarrow\lim_{t\to\infty} V_i\left(\mathbf{Y},\mathbf{x}_t,t\right)=\delta_i\left(\mathbf{x},\mathbf{Y}\right) \tag{2.127}$$

$$\Longrightarrow \lim_{t\to\infty}\xi_{FLVQ}\left(t\right)\approx\xi_E^2 \tag{2.128}$$

La borrosidad inicial debida a grandes valores de $m\left(t\right)$ se encargan de mover los vectores de código a la región ocupada por los datos de la muestra. La programación de $m\left(t\right)$ es crítica como ya se ha dicho.

2.7.Agrupamiento no estacionario y cuantización vectorial adaptativa

En esta sección primeramente damos una definición de la cuantización vectorial adaptativa. Relacionamos un caso especial de cuantización vectorial adaptativa con el agrupamiento no estacionario. Finalmente, discutimos la aplicación apropiada de las redes neuronales competitivas como mecanismos de cuantización vectorial adaptativa para resolver problemas de agrupamiento no estacionario.

2.7.1.Cuantización vectorial adaptativa

En la sección 2.5 hemos revisado las definiciones de la cuantización vectorial y el problema del agrupamiento en el caso estacionario. El proceso estocástico $\{X\left(t_j\right);j=0,1,2,...\}$ que modela la fuente de datos era una secuencia de vectores aleatorios i.i.d. para los que la función de densidad de probabilidad es invariante en el tiempo

$$p\left(\mathbf{x},t_j\right) = p\left(\mathbf{x}\right), \tag{2.129}$$

$$p\left(\mathbf{x}\left(t_j\right),...,\mathbf{x}\left(t_{j-m}\right)\right) = \prod_{n=0}^{m} p\left(\mathbf{x}\right);\forall m>0$$

donde $p(\mathbf{x},t_j)$ es la d.d.p. de los datos en el instante t_j. Incluso cuando los vectores aleatorios no son independientes, el proceso estocástico permanece estacionario cuando la función de densidad de probabilidad conjunta (f.d.p.c.) es invariante en el tiempo

$$p(\mathbf{x}(t_j),...,\mathbf{x}(t_{j-m})) = p(\mathbf{x}(t_i),...,\mathbf{x}(t_{i-m}));t_i \neq t_j \tag{2.130}$$

donde m es la extensión de la memoria del proceso. Si el proceso estacionario es gausiano, la media y los momentos de segundo orden son suficientes para caracterizarlo como estacionario. El proceso es estacionario en el sentido de la media si su media permanece constante

$$E_{\mathbf{x}}[\mathbf{x}(t)] = \mu; \forall t > 0 \tag{2.131}$$

El proceso es estacionario en relación a las correlaciónes si estas son invariantes en el tiempo

$$E_{\mathbf{x}}\left[\mathbf{x}(t_j)\,\mathbf{x}^t(t_{j-m})\right] = \mathbf{R}_m; m \geq 0 \tag{2.132a}$$

Si el proceso es estacionario, podemos poner el problema de la construcción de un cuantificador vectorial en el marco de la sección 2.5 si postulamos un predictor

$$\widetilde{\mathbf{x}}(t_j) = P(\mathbf{x}(t_{j-1}),...,\mathbf{x}(t_{j-m})) \tag{2.133}$$

cuya expresión depende de la forma de la función de densidad de probabilidad conjunta invariante en el tiempo como se define en la ecuación 2.130. El error de predicción

$$\mathbf{e}(t_j) = \mathbf{x}(t_j) - \widetilde{\mathbf{x}}(t_j) \tag{2.134}$$

es una secuencia de vectores aleatorios i.i.d. para los que se puede diseñar un cuantificador vectorial $\widehat{\mathbf{e}} = \mathcal{Q}(\mathbf{e})$. La reconstrucción de la señal se realiza añadiendo el error de cuantización a la señal predicha

$$\widehat{\mathbf{x}}(t_j) = \widehat{\mathbf{e}}(t_j) + \widetilde{\mathbf{x}}(t_j) \tag{2.135}$$

Los cuantizadores vectoriales predictivos (PVQ) discutidos en [42] son de hecho extensiones vectoriales de las técnicas de codificación escalar predictiva basadas en modelos predictivos lineales. El predictor de estados finitos también discutido en [42] assume dependencias temporales más complejas que pueden ser modeladas por una máquina de estados finitos. La definición y exploración de VQ predictivo no lineal es un área de trabajo inexplorada hasta donde conocemos.

En el caso no estacionario, la f.d.p.c no invariante en el tiempo. Esto es, la ecuación 2.130 no se cumple para todo los instantes de tiempo. El proceso no estacionario más sencillo es el paseo aleatorio (*random walk*) [91] o proceso de Wiener, un proceso autoregresivo inestable de varianza no acotada, que se utiliza a veces como *benchmark* [35]. Podemos caracterizar un proceso como localmente estacionario cuando satisface

$$p\left(\mathbf{x}\left(t_j\right),...,\mathbf{x}\left(t_{j-m}\right)\right) = p\left(\mathbf{x}\left(t_i\right),...,\mathbf{x}\left(t_{i-m}\right)\right); t_i < t_{j+Nm}; N >> 0 \qquad (2.136)$$

lo que significa que la ecuación 2.130 se cumple para periodos de tiempo finitos pero lo bastante largos. Para procesos localmente estacionarios podemos postular un predictor dependiente en tiempo

$$\widetilde{\mathbf{x}}\left(t_j\right) = P_{t_j}\left(\mathbf{x}\left(t_{j-1}\right),...,\mathbf{x}\left(t_{j-m}\right)\right) \qquad (2.137)$$

que será estimado a intervalos regulares. El error de predicción sigue siendo una fuente de vectores aleatorios i.i.d. La aproximación predictiva puede extenderse a los procesos no estacionarios asumiendo el costo de la repetición de los procesos de identificación y estimación de los parámetros del modelo predictivo.

La aproximación adaptativa intenta evitar estas deventajas de los cuantizadores predictivos mediante la modificación online del cuantizador vectorial. Es importante notar que la asunción de estacionariedad local sigue siendo importante para la aproximación adaptativa, porque implica que hay suficientes datos disponibles para la adaptación. La VQ adaptativa se discute en [42]. Aproximaciones básicas son la reducción de la media, la adaptación de la ganancia y algunas estrategias para el relleno del libro de códigos[35],[154],[16]. También se han propuesto redes neuronales competitivas como herramientas computacionales para realizar la VQ adaptativa. En [42] se declara que son de poca utilidad debido a sus largos tiempos de convergencia y falta de robustez. En [87] se prueban nuevamente para secuencias de imágenes con cierto éxito debido sobre todo a la selección cuidadosa de las condiciones iniciales. De hecho, la mayor parte de las aplicaciones de redes neuronales competitivas a la construcción de cuantizadores vectoriales se han realizado sobre imágenes fijas y secuencias de imágenes con poca variabilidad (i.e.: caras parlantes).

La aproximación adaptativa produce un VQ variante en el tiempo

$$\widehat{\mathbf{x}}\left(t\right) = \mathcal{D}_t\left(\mathcal{E}_t\left(\mathbf{x}\right)\right) \qquad (2.138)$$

cuya calidad se mide mediante la esperanza del error variante en el tiempo

$$\xi\left(t\right) = E_{\mathbf{x},t}\left[\varepsilon\left(\mathbf{x},\widehat{\mathbf{x}}\left(t\right)\right)\right] = \int \varepsilon\left(\mathbf{x},\widehat{\mathbf{x}}\left(t\right)\right) p\left(\mathbf{x},t\right) d\mathbf{x} \qquad (2.139)$$

2.7.2.Muestreo y estacionariedad local

El proceso estocástico vectorial $\{\mathbf{x}(t)\}$ que está siendo cuantizado proviene de un procedimiento de muestreo que puede influenciar su naturaleza probabilística. Cuando tratamos con procesos escalares, como señales acústicas o series temporales financieras, los procesos estocástico vectoriales se construyen mediante la agregación de las muestras escalares en un vector [42], [91] Sea $\{x(t_0 + n\triangle t)\}$ la señal escalar muestreada, el proceso vectorial de dimensión d se construye como se especifica en la siguiente ecuación

$$\mathbf{x}(t) = (x(t-(d-1)\triangle t), ..., x(t)); t = t_0 + Nd\triangle t; N = 1, 2, ... \qquad (2.140)$$

Considerese el caso en que el proceso original muestra periodicidad de periodo d, el proceso vectorial resultante de la agregación $\{\mathbf{x}(t)\}$ no será periódico. Otros procedimientos de muestreo, como la extracción de los coeficientes mel-cepstrum para tareas de reconocimiento de voz son también dependientes críticamente de la ventana temporal usada para los cálculos.

En el caso no estacionario general, la evolución de los procesos físicos subyacentes es arbitraria tanto en la forma funcional de la f.d.p.c. como en la velocidad de los cambios. Si los procedimientos de muestreo son lentos en relación a los cambios de estos procesos, la muestra estará compuesta de una secuencia de vectores aleatorios independientes que obedecen distintas distribuciones de probabilidad. Entonces, una muestra obtenida en una ventana temporal (t_i, t_f) debe denotarse como

$$\aleph(t_i, t_f) = \{\mathbf{x}(t_1), ..., \mathbf{x}(t_n)\} \quad \text{con} \quad t_1 = t_i; t_f = t_n \qquad (2.141)$$

donde $\mathbf{x}(t_j)$ es una muestra de la señal aleatoria en el instante t_j cuya densidad de probabilidad se denota $p(\mathbf{x},t_j)$. La mayor parte de las instancias de VQ adaptativo en la literatura aplicadas al proceso de la imagen asumen que una imagen fija obedece esta caracterización, e intentan realizar adaptación dentro de la imagen (*intraframe*). Desde nuestro puento de vista, es difícil verificar alguna forma de estacionariedad local en la muestra descrita en la ecuación 2.141. Por tanto, el ajuste de los algoritmos es una tarea delicada, y los resultados informados en la literatura deberían ser tomados con precaución. Una decisión clave en nuestro trabajo es asumir un marco de muestreo distinto. Podemos asumir que el proceso de muestreo es rápido en relación a los cambios físicos subyacentes. Por tanto, las muestras obtenidas en una ventana temporal se pueden considerar como un conjunto de vectores aleatorios i.i.d. El procedimiento de muestreo produce una

secuencia de estos conjuntos, que denotamos como

$$\aleph(t) = \{\mathbf{x}_1(t), ..., \mathbf{x}_n(t)\} \quad t = 0, 1, 2, ... \tag{2.142}$$

donde la d.d.p. $p(\mathbf{x},t)$ permanece invariante para la ventana temporal en la que se obtiene la muestra. Esta definición es extraña para el proceso de señal 1D, pero es muy natural para secuencias de imágenes. De hecho, asumimos que t es el número de la imagen en la secuencia. Notese que tenemos el mismo tamaño de la muestra n para todos los instantes t. El tamaño de la muestra no sólo depende del tamaño de la imagen, puede estar relacionado con características específicas de los algoritmos. Cada imagen es un proceso localmente estacionario que permanece invariante durante el tiempo suficiente para permitir la adaptación.

2.7.3.VQ adaptativo y agrupamiento no estacionario

Tradicionalmente, el VQ adaptativo para la compresión de señal con pérdida se analiza en el marco de la teoría de ratio-distorsión [5]. La optimalidad de los algoritmos de AVQ consiste en su habilidad para alcanzar la función teórica de ratio-distorsión (si se conoce). Por ejemplo, [154] prueba que el algoritmo *gold-washing* es óptimo en este sentido para fuentes i.i.d que satisfacen algunas condiciones naturales. Sin embargo, este resultado tiene poco que decir de su rendimiento como un algoritmo AVQ en un entorno no estacionario. La mayor parte de los algoritmos AVQ se prueban empíricamente sobre procesos (imágenes) cuya función ratio-distorsión es desconocida. Finalmente, la mayor parte de los autores interpretan los algoritmos AVQ como procedimientos de selección que extraen los libros de códigos instantaneos de un VQ Universal abstracto que cuantiza optimamente la señal. Esta interpretación es natural para algoritmos de relleno de los libros de códigos, pero menos clara para otras como las redes neuronales competitivas.

Tomando en cuenta que VQ (y AVQ) puede ser usada para otras tareas distintas de la codificación de señal, como la segmentación de imágenes [71], [72], [95], [109], [137], [143] o métodos de proyección no lineal [75]. Proponemos definir AVQ como un proceso de minimización dinámica de la medida de calidad de los cuantizadores vectoriales variando en el tiempo

$$\min_{\{\mathcal{Q}_t\}} \sum_{t \geq 0} \xi(t), \tag{2.143}$$

donde $\{\mathcal{Q}_t\}$ es la secuencia de funciones de codificación producidas por la adaptación. Para cuantificadores del tipo del vecino más cercano, el AVQ se define como la

búsqueda de la secuencia de libros de códigos

$$\mathbf{Y}(t) = \{\mathbf{y}_1(t) \ , .., \mathbf{y}_c(t)\}; t = 0, 1, 2, ... \tag{2.144}$$

que minimiza la distorsión de la sequencia de cuantizaciones realizadas usandolos en cada instante de tiempo. Cuando el procedimiento de muestreo produce una secuencia de muestras i.i.d. como en la ecuación 2.142, y la medida de distorsión es la distancia euclídea al cuadrado, el AVQ puede ser formulado como el siguiente problema de minimización:

$$\begin{aligned} \min_{\{\mathcal{Q}_t\}} \sum_{t\geq 0} \widehat{\xi_E^2}(t) &= \min_{\{\mathbf{Y}(t)\}} \sum_{t\geq 0} \sum_{j=1}^{n} \sum_{i=1}^{c} \|\mathbf{x}_j(t) - \mathbf{y}_i(t)\|^2 \delta_{ij}(t) \qquad (2.145) \\ \delta_{ij}(t) &= \begin{cases} 1 & i = \arg\min_{k=1,..,c} \left\{\|\mathbf{x}_j(t) - \mathbf{y}_k(t)\|^2\right\} \\ 0 & \text{sino} \end{cases} \end{aligned}$$

Para la definición de agrupamiento no estacionario necesitamos asumir una secuencia de muestras de propiedades estadísticas cambiantes, como en la ecuación 2.142. Entonces, el agrupamiento no estacionario trata de obtener en cada instante de tiempo la partición de la muestra

$$P(\aleph(t)) = \{\aleph_1(t), .., \aleph_c(t)\} \tag{2.146}$$

que es óptima en el sentido especificado por alguna función objetivo que cuantifica la calidad del agrupamiento realizado. Cuando la función criterio del agrupamiento es la varianza intra-grupo, y las muestras de los datos vienen de un proceso de tiempo discreto muestreado igualmente en cada instante de tiempo, el agrupamiento no estacionario es el mismo problema que el VQ adaptativo definido en la ecuación 2.145.

2.7.4. Applicación de las redes neuronales competitivas como AVQ para agrupamiento no estacionario

El problema AVQ tal y como se formula en la ecuación 2.145 es un problema de programación dinámica de horizonte temporal infinito. Este problema se puede hacer más tratable asumiendo

$$\min_{\{\mathcal{Q}_t\}} \sum_{t\geq 0} \xi(t) = \sum_{t\geq 0} \min_{\mathcal{Q}_t} \xi(t) \tag{2.147}$$

que viene a ser para una muestra dada, y la cuantización por el vecino más cercano en distancia euclídea

$$\min_{\{\mathbf{Y}(t)\}} \sum_{t\geq 0} \widehat{\xi_E^2}(t) = \sum_{t\geq 0} \min_{\mathbf{Y}(t)} \widehat{\xi_E^2}(t) \quad (2.148)$$

$$= \sum_{t\geq 0} \min_{\mathbf{Y}(t)} \sum_{j=1}^{n} \sum_{i=1}^{c} \|\mathbf{x}_j(t) - \mathbf{y}_i(t)\|^2 \delta_{ij}(t) \quad (2.149)$$

La minimización de la secuencia de funciones de error dependientes del tiempo puede ser realizada de forma independiente en cada paso de tiempo. Esta es una asunción razonable para el proceso de la imagen. Cuando se aplica a las imágenes corresponde a la codificación intra-trama. Cuando se aplica a las diferencias entre imágenes o a los campos vectoriales de movimiento, corresponde a la cuantización para codificación entre tramas.

Además, asumiendo que la estadísticas de las muestras cambian suavemente, se puede asumir la variación suave y acotada de los libros de códigos óptimos en instantes de tiempo sucesivos. Por tanto, el libro de códigos óptimo encontrado en el paso de tiempo anterior se puede considerar como una buena condición inicial para la búsqueda del libro de códigos óptimo en el instante actual.

La aplicación adaptativa de los algoritmos de redes neuronales se hace como sigue: En el instante t los representantes iniciales de los agrupamientos son los calculados a partir de la muestra obtenida en el instante $t-1$. Los vectores de la muestra $\aleph(t) = \{\mathbf{x}_1(t), .., \mathbf{x}_n(t)\}$ se presentan secuencialmente y aleatoriamente como las entradas para calcular las ecuaciones de adaptación y obtener un nuevo conjunto de representantes de los agrupamientos. Una característica distintiva de los trabajos experimentales presentados en los siguientes capítulos es que imponemos una adaptación en un solo paso sobre la muestra.

2.8. Conclusiones

Hemos comenzado retomando la relación entre Agrupamiento y Cuantización Vectorial en el caso estacionario. Esta relación se extiende al caso no estacionario, relacionando Cuantización Vectorial Adaptativa y Agrupamiento No Estacionario. La definición de los algoritmos de redes neuronales competitivas como la minimización mediante descenso de gradiente estocástico, de funciones objetivo dadas (o sospechadas) no lleva a la consideración de los algoritmos SOM, FLVQ y SCS

como procesos de minimización cuyo objetivo final es la minimización de la distorsión o error cuadrático medio. Estos algoritmos realizan esto mediante la minimización de una secuencia de funciones objetivo que convergen funcionalmente a la distorsión. Este proceso se dirige a obtener mayor robustez frente a las condiciones iniciales y un rendimiento mejorado comparado con el SCL desnudo. Discutimos estas convergencias funcionales para cada algoritmo, identificando los parámetros de las funciones de vecindad que la controlan. Proponemos una programación exponencial de estos parámetros que resulta en una adaptación rápida, mejora del rendimiento y robustez. Estas propiedades permiten la aplicación de SOM, FLVQ y SCS como algoritmos de AVQ a tareas de Agrupamiento No Estacionario, por lo que los probaremos sobre la cuantización del color de secuencias de imágenes en el siguiente capítulo.

3. Convergencia en un solo paso sobre la muestra: estudio empírico

En las aplicaciones descritas en los capítulos posteriores, necesitamos que la respuesta sea calculable en un tiempo no excesivo. En el caso de las imágenes de resonancia magnética, estos tiempos son del orden de segundos, mientras que en caso del cálculo del flujo óptico estos tiempos son del orden de fracciones de segundo. De todas maneras, las secuencias de entrenamiento habituales para las redes neuronales competitivas son excesivamente lentas. Es por ello que probamos con secuencias rápidas de entrenamiento. En este capítulo exploramos la calidad que tienen los algoritmos competitivos cuando el entrenamiento se realiza en un sólo paso sobre la muestra de datos. En primer lugar tratamos con un conjunto de datos bidimensionales sintéticos. Más adelante consideramos como marco de trabajo una instancia del agrupamiento no estacionario: la cuantización de color no estacionaria sobre secuencias de imágenes [1]. El caracter no estacionario de los datos viene de la impredecibilidad de la distribución de los píxeles en las imágenes. Concretamente, testeamos el rendimiento del SOM, SCS y FLVQ sobre esta tarea [49], [50].

[1] En la cuantización del color de imágenes, la medida de la calidad de los resultados del agrupamiento es la distorsión producida por la sustitución de los colores auténticos de los píxeles por el representante de color más cercano. En la medida en que nuestro interés principal reside en los resultados comparativos de los distintos algoritmos para el cálculo del cuantizador de color, trabajaremos en el espacio RGB y usaremos la distancia euclídea como la medida de similitud. Los algoritmos de cuantización de color más extendidos, basados en las ideas originales de Heckbert [61], trabajan en el espacio RGB. Sin embargo, son necesarias algunas palabras de justificación. En primer lugar, somos conscientes de que la distancia euclídea no es una distancia perceptivamente consistente en el cubo RGB: puntos cercanos en el cubo RGB pueden ser percibidos como colores muy distintos, y viceversa. Otros espacios de color [116] como el UVW, se consideran más cercanos a reflejar la distancia perceptual entre colores. Sin embargo, resultados experimentales citados en [116] (pp.167) sugieren que la realización de la cuantización del color en el cubo RGB no introduce una gran pérdida perceptiva. Esto puede ser una justificación para muchos de los algorimos en la literatura que trabajan en este espacio [148] [69] [90] [40] [107] [137] [144] [149].

En la sección 3.1 se prentan las secuenicas de valores de los parámetros que controla el proceso de aprendizaje. La sección 3.2 presenta los resutados sobre los datos estacionarios. La sección 3.3 presenta los resultados sobre los datos de la cuantización del color. Finalmente, la sección 3.4 presenta las conclusiones del capitulo.

3.1.Ajuste de los parámetros de control

En esta sección comentaremos el ajuste de los parámetros de control críticos de las redes neuronales competitivas cuya convergencia pretendemos estudiar. Primero discutiremos la programación de la velocidad de aprendizaje, puesto que es común a todos los algoritmos. Después dicutiremos la programación de cada uno de los parámetros de vecindad que controla la convergencia funcional a la regla de aprendizaje competitivo simple.

El ajuste de los parámetros de control no es trivial puesto que puede condicionar incluso el caracter del algoritmo. Por ejemplo,ha sido demostrado que el SOM tiene convergencia débil a estados organizados en el caso de espacios multidimensionales [34]. Todas las pruebas descansasn en la definición de un SOM con velocidad de aprendizaje constante y función de vecindad invariante en el tiempo. También, para sistemas adaptativos que deben permanecer flexibles para seguir los cambios ambientales, la velocidad de aprendizaje no debe hacerse cero o debería reinicializarse cuando se detectan o sospechan cambios en el ambiente. Por tanto, no es inhabitual encontrar velocidades de aprendizaje que aseguran que la adaptación no se congelará en el tiempo [145] [139] [38]. Sin embargo, esta velocidad debe ser muy pequeña para prevenir inestabilidades, y tales pequeñas velocidades producen convergencia muy lenta.

En la proposición original de SCS en [153], la programación de la velocidad de aprendizaje se reinicializa a intervalos que crecen cuadráticamente, y hay algunos parámetros que deben ser ajustados finamente para obtener resultados razonables. La reinicialización de la velociadad de aprendizaje está enlazada con la idea de que el aprendizaje es una secuencia de minimizaciones a temperaturas decrecientes, porque el proceso completo se presenta como un algoritmo de enfriamiento estadístico. Algunos estudios [101] tratan de asegurar que la secuencia de valores de velocidades de aprendizaje es óptima en el sentido de que todas la presentaciones de datos input tienen la misma contribución al estado final. Estos estudios no muestran como aliviar el costo computacional (pueden incluso involucrar un gran número de repeticiones del proceso de aprendizaje completo para refinar la

secuencia de valores de velocidades de aprendizaje), y no mejora la convergencia a los mínimos de la función de energía, porque están dirigidos a proporcionar invarianza al orden de presentación del input.

Asumimos que las redes competitivas son un método de descenso de gradiente estocástico de una función objetivo dada. Para garantizar la convergencia teórica a un mínimo local de esta función, la velocidad de aprendizaje debe cumplir las condiciones de Robins-Monro ([39], [79], [85], [118], [135]):

$$\begin{aligned} \lim_{\tau \to \infty} \alpha(t) &= 0 \\ \sum_{\tau=0}^{\infty} \alpha(t) &= \infty \\ \sum_{\tau=0}^{\infty} \alpha^2(t) &< \infty \end{aligned} \tag{3.1}$$

Estas condiciones implican procesos computacionales muy largos.

Contrariamente a lo expresado en [8] [153] consideramos la velocidad de aprendizaje como un parámetro de control del proceso numérico de descenso estocástico del gradiente mientras que la función de vecindad es característica de la función objetivo que es minimizada, por lo que no las consideramos agregadas en un único parámetro. Como ya hemos dicho, nuestro interes es probar la mejora dada por la consideración de una función objetivo que da lugar a una función de vecindad particular. La programación de la velocidad de aprendizaje debe ser igualmente aplicable a todos los algoritmos. La que hemos usado en los experimentos de este capítulo tiene las siguientes características.

1. Asumimos un proceso de adaptación en un paso sobre la muestra. Esto es, la muestra se presenta sólo una vez para realizar la adaptación.

2. Las unidades tienen velocidad de aprendizaje local $\alpha_i(t)$ $i = 1, .., c$. Esto puede interpretarse como la realización simultánea de tantos descensos de gradiente como unidades. Hemos encontrado que esta estrategia mejora la convergencia global y reduce el riesgo de que las unidades se queden congeladas debido a un orden desventajoso en la presentación de los datos.

3. Progresión geométrica de la velocidad de aprendizaje

$$\alpha_i(t) = 0.1\,(1 - t_i/n) \tag{3.2}$$

donde n es el tamaño de la muestra, y t_i es el tiempo local de adaptación cuyo valor exacto depende de la regla de aprendizaje. En el caso del SCL se

calcula como

$$t_i = \sum_{k=1}^{t} \delta_i \left(\mathbf{x}\left(k\right), \mathbf{Y}\left(k\right) \right) \tag{3.3}$$

y en el caso general se calcula como la acumulación de los valores de la función de vecindad durante el proceso de aprendizaje, lo que corresponde al número de adaptaciónes sufridas por la unidad.

$$t_i = \sum_{k=1}^{t} V_i \left(\mathbf{x}\left(k\right), \mathbf{Y}\left(k\right) \right) \tag{3.4}$$

Esta definición no cumple las condiciones de convergencia dadas en las ecuaciones 3.1: la velocidad de aprendizaje local alcanza el valor cero sólo cuando todas la muestra ha sido monopolizada por la misma unidad, mientras que la segunda condición no se cumple nunca. Sin embargo, hemos encontrado que es muy conveniente para adaptaciónes en un solo paso sobre la muestra como pretendemos realizar en nuestro trabajo.

Discutimos a continuación la programación de las funciones de vecindad $V_i\left(\mathbf{x}, \mathbf{Y}, \tau\right)$. Como están relacionadas con la función que se minimiza se gestionan como características globales de los algoritmos cuya variación no depende de la unidad ganadora. Hemos escogido un decrecimiento exponencial a la función vecindad nula. La velocidad de convergencia a esa función se denota r. Esto es, asumimos que para el SOM, FLVQ y SCS

$$V_i\left(\mathbf{x}, \mathbf{Y}, t\right) = \delta_i\left(\mathbf{x}, \mathbf{Y}\right) \quad t \geq \frac{n}{r}, \tag{3.5}$$

donde n es el tamaño de la muestra, y t es el tiempo de adaptación. Cuando $r = 1$ el vecindario no llega a ser nulo en todo el entrenamiento: no hay fase SCL y la convergencia al SCL no tiene efecto. Cuando r crece la convergencia a SCL es más rápida y la fase SCL es más larga. No consideramos $r < 1$ puesto que no tienen ningún sentido en una estrategia de adaptación en un paso sobre la muestra.

En el caso del SOM hemos asumido una topología 1D de los índices de las unidades de la red. El radio inicial del vecindario es $v\left(0\right) = v_0$. El tamaño del vecindario está dato por $v\left(t\right)$ que tienen la siguiente expresión

$$v\left(t\right) = \left\lceil \left(v_0 + 1\right)^{\left(1 - \frac{r}{n} t\right)} \right\rceil - 1 \quad t < \frac{n}{r}. \tag{3.6}$$

En el caso de FLVQ la extensión de la influencia de las adaptaciónes se controla mediante el exponente $m\left(\tau\right)$. Para propósitos prácticos, un valor de $m_f = 1.1$ hace que la regla de FLVQ sea idéntica a SCL, y debe ser alcanzada en $t = \frac{n}{r}$.

Por tanto $m\left(\frac{n}{r}\right) = m_f = 1.1$. El valor inicial del exponente es $m\left(0\right) = m_0$. La evolución del exponente se controla por la siguiente función:

$$m\left(t\right) = m_0 \left(\frac{m_f}{m_0}\right)^{\frac{r}{n}t} \quad t < \frac{n}{r}. \tag{3.7}$$

En el caso de SCS, el parámetro que controla la entensión de la adaptación inducida por un input es la desviación estandard asumida $\sigma\left(t\right)$. El valor inicial de la desviación estandard es $\sigma\left(0\right) = \sigma_0$. También en este caso hemos aplicado un descenso exponencial de este parámetro:

$$\sigma\left(t\right) = \left(\sigma_0 + 1\right)^{\left(1-\frac{r}{n}t\right)} - 1 \quad t < \frac{n}{r} \tag{3.8}$$

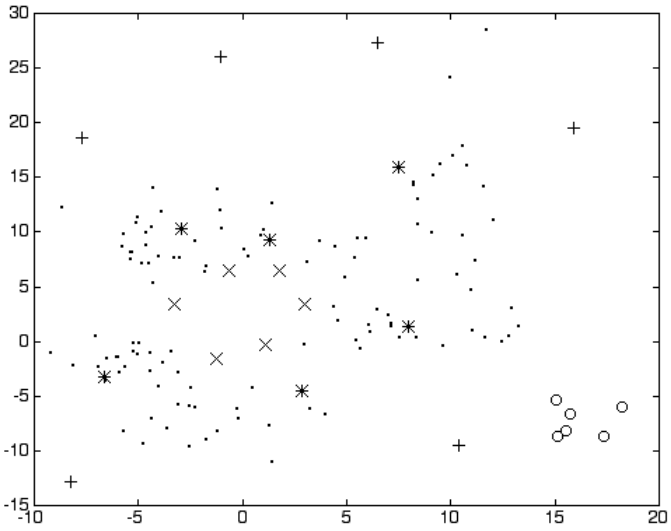
El rendimiento de SCS es muy sensible al valor inicial de la desviación estandard σ_0. Valores muy altos fuerzan la convergencia a la media de la muestra, mientras que valores pequeños producen una convergencia prematura a SCL. Su interpretación estadística permite una aproximación más sistemática que la estrategia de prueba y error para fijar σ_0. Hemos denotado $\widehat{\sigma}_{i,0}$ los estimadores locales de las desviaciónes estandard basados en las distancias entre vectores código. Estos estimadores se calculan para obtener los máximos intervalos de confianza del 95% no solapados bajo la asunción de que los vectores codigo son los centros de d.d.p. gausianas isotrópicas [54].

3.2.Resultados experimentales sobre datos estacionarios

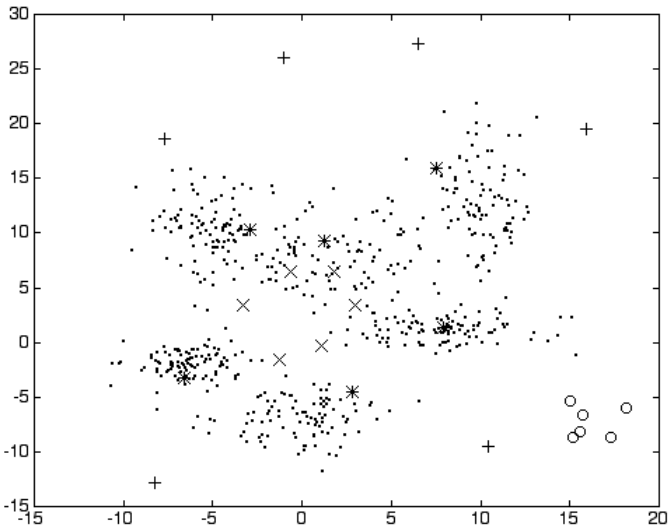
Hemos realizado algunos experimentos preliminares de agrupamiento sobre datos estacionarios para explorar la mejora que SOM, FLVQ y SCS obtienen sobre SCL cuando se aplican como procedimientos de inicialización robustos. Los resultados se comparan también con los obtenidos por el algoritmo Isodata. Se escogen cuatro libros de códigos como representantes de muy diferentes condiciones iniciales, de forma que los experimentos muestran una exploración selectiva y cualificada de la robustez bajo diversas (buenas y malas) condiciones iniciales.

3.2.1.Los datos estacionarios

Los datos estacionario corresponden a muestras 2D generadas a partir de una mezcla de 6 distribuciones gausianas con distintas varianzas. Hemos asumido $c = 6$ para todos los algoritmos. En la figura 3.1 mostramos las muestras que



(a)



(b)

Figura 3.1: Muestras de datos para los experimentos sobre agrupamiento con datos estacionarios mostrando los libros de códigos iniciales usados en los experimentos (a) muestra con 120 puntos, (b) muestra con 600 puntos

Tabla 3.1: Resultados del algoritmo de las k-medias en los datos y libros de códigos de la figura 3.1. Las distorsiónes iniciales antes de la aplicación del k-medias, distorsiónes obtenidas por los libros de códigos del K-medias, Número de iteraciones necesarias para alcanzar los libros de códigos finales

distortion	Sample 1			Sample 2		
	Initial	Isodata	iter.	Initial	Isodata	iter.
Codebook1	14201	1794	5	72108	8703	25
Codebook2	5922	1525	7	27930	6993	10
Codebook3	45079	4056	6	220839	8104	34
Codebook4	1772	1294	5	8481	5472	3

Tabla 3.2: Resultados de la adaptación en un paso con SCL. Distorsión de las muestras cuantizadas usando los libros de códigos calculados por el SCL en un solo paso sobre la muestra

distortion	Sample 1	Sample 2
Codebook1	2350	13403
Codebook2	2072	10143
Codebook3	12200	45412
Codebook4	1454	7227

hemos usado y los libros de códigos usados en los experimentos (denotados por "o"," x","+" y "*"). La segunda muestra es diez veces mayor que la segunda. La elección de los libros de códigos iniciales siguen los siguientes criterios:

1. Codebook1 ("o"): es una condición inicial fuera de la región del espacio ocupada por la muestra, con todos los libros de códigos cercanos entre sí. Esta situación produce facilmente vectores de código congelados (sin muestras asociadas). Podemos considerla como una mala condición inicial.

2. Codebook2 ("x"): es una buena condición inicial, con los vectores de código situados en un hueco central de la región ocupada por las muestras.

3. Codebook3 ("+"): es una condición inicial fuera de la muestra, alrededor de ella, menos sujeta a que se produzcan vectores congelados.

4. Codebook4 ("*"): es la mejor condición inicial, con los vectores código distribuidos regularmente en la región de la muestra.

3.2.2.El algoritmo básico de referencia: K-medias

En adelante nos referimos indistintamente al algoritmo K-medias como Isodata. En la tabla 3.1 presentamos resultados sobre las muestras de datos de la aplicación del algoritmo K-medias para buscar los representantes de los agrupamientos. El criterio de parada prueba que la diferencia absoluta entre las distorsiónes de iteraciones consecutivas es menor que 0.01 (lo que es equivalente a una variación relativa de 10^{-6} de la distorsión relativa). De hecho el algoritmo k-medias se detiene cuando no hay cambio en el libro de códigos. Mostramos en la tabla 3.1, para cada muestra y condición inicial, la distorsión inicial para cada libro de códigos inicial, la distorsión tras la aplicación del Isodata y el número de iteraciones que necesita para alcanzar el mínimo local. Las distorsiónes finales confirman nuestra calificación de los libros de códigos iniciales. Los codebooks 2 y 4 son buenas condiciones iniciales que conducen a los mejores óptimos locales, el codebook 4 es el mejor. Los codebooks 1 y 3 son malas condiciones iniciales. Aunque el algoritmo K-medias los mejora considerablemente (90% de reducción de la distorsión) los mínimos locales alcanzados a partir de ellos son peores que aquellos alcanzados desde los codebooks 2 y 4. La peor condición inicial es el codebook 3. El número de iteraciones para alcanzar el mínimo local aumenta con el tamaño de la muestra, pero también con la "'maldad" del libro de códigos inicial. El efecto de las condiciones iniciales en el número de iteraciones es menos claro en la muestra pequeña que en la grande: malas condiciones iniciales implican cálculos más largos. Esto es de alguna importancia debido a que pedimos a las aproximaciones de redes neuronales competitivas que realizen sólo una iteración sobre la muestra.

3.2.3.Resultados del algoritmo competitivo simple (SCL)

En la tabla 3.2 se muestran los resultados del entrenamiento del SCL con un sólo paso sobre la muestra. La programación de la velocidad de aprendizaje es la discutida en la sección previa. Como es de esperar, los resultados del algortimo K-medias son mejores que los del SCL en un único paso sobre la muestra. Sin embargo, el entrenamiento en un paso sobre la muestra de SCL da una mejora significativa de los libros de códigos iniciales. Los buenos libros de códigos son mejorados en un grado menor. No tratamos de mostrar que SCL es superior al algoritmo K-medias. El objetivo es mostrar que las mejoras introducidas por

nuestras programaciones de las funciones de vecindad hacen que las redes competitivas se acerquen a los resultados del algoritmo K-medias pero en un solo paso de adaptación sobre la muestra.

3.2.4.Resultados de SOM, FLVQ y SCS

En primer lugar exploramos la sensibilidad del SOM, FLVQ y SCS a la programación de los parámetros de control de sus funciones de vecindad. Para ello, hemos realizado la adaptación en un paso para cada libro de códigos inicial, tamaño de muestra y diferentes combinaciones de r la velocidad de convergencia a SCL. Este experimento de sensibilidad está dirigido a ilustrar el efecto de estos parámetros. En general los algoritmos mejoran al SCL. Realizan, como se pretende demostrar, inicializaciones robustas. En algunos casos incluso mejoran al algoritmo Isodata a pesar de que sólo realizan un paso sobre la muestra. Los resultados son similares en todas las pruebas que hemos realizado con los datos usados aquí y con otros conjuntos de datos, incluso con los datos no estacionarios como se verá más tarde.

Tabla 3.3: Los mejores resultados de SOM, FLVQ y SCS en la adaptación en un paso sobre las muestras de la figura 3.1 para cada uno de los libros de códigos mostrados en ella. Destacamos los resultados que mejoran a los dados para el algoritmo de las k-medias en la tabla 3.1.

distortion	Sample 1			Sample 2		
	SOM	FLVQ	SCS	SOM	FLVQ	SCS
Codebook1	1970	2302	2071	**7692**	13339	13406
Codebook2	1613	2073	1895	8137	8554	8376
Codebook3	**2061**	6240	5372	**6559**	15028	11180
Codebook4	1365	1431	1330	8037	6865	6858

Los resultados detallados de esta exploración de la sensibilidad de los algoritmos sobre la muestra pequeña se presentan en las figuras 3.2, 3.2.4 y 3.2.4 para el SOM, el FLVQ y el SCS respectivamente. Cada gráfica en estas figuras muestra la distorsión que se alcanza comenzando de alguno de los libros de códigos iniciales marcados en la figura 3.1. Cada punto en las superficies corresponde a una combinación de valores de los parámetros velocidad de convergencia a SCL y el

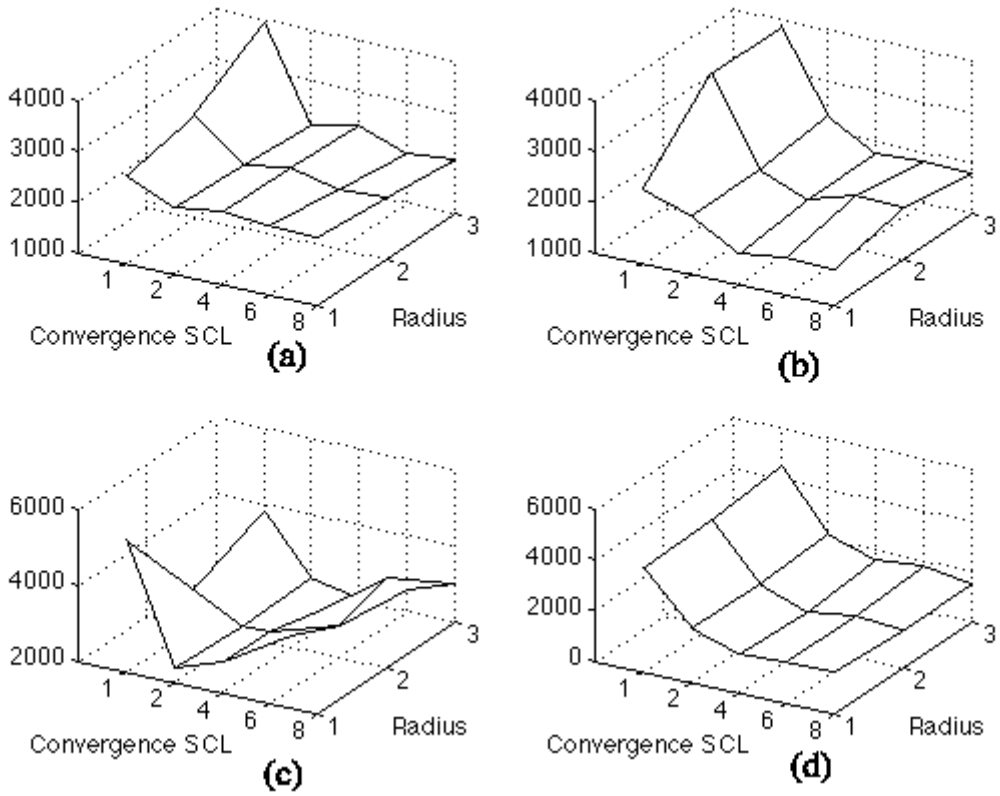


Figura 3.2: Exploración de la sensibilidad del SOM al radio del vecindario ($v_0 = 1, 2, 3$) y a la velocidad de convergencia al SCL ($r = 1, 2, 4, 6, 8$) . Resultados obtenidos sobre la muestra pequeña comenzando libros de códigos identificados en la figura 3.1

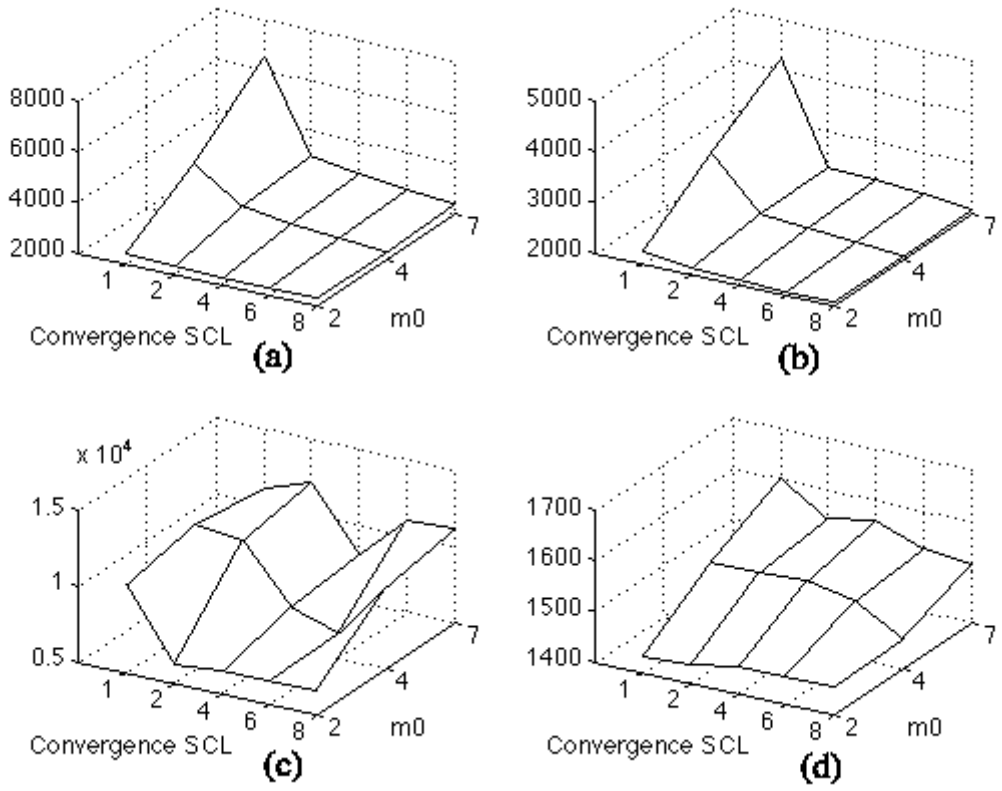


Figura 3.3: Exploración de la sensibilidad del FLVQ al exponente inicial ($m_0 = 2, 4, 7$) y a la velocidad de convergencia al SCL ($r = 1, 2, 4, 6, 8$) . Resultados obtenidos sobre la muestra pequeña comenzando libros de códigos identificados en la figura 3.1

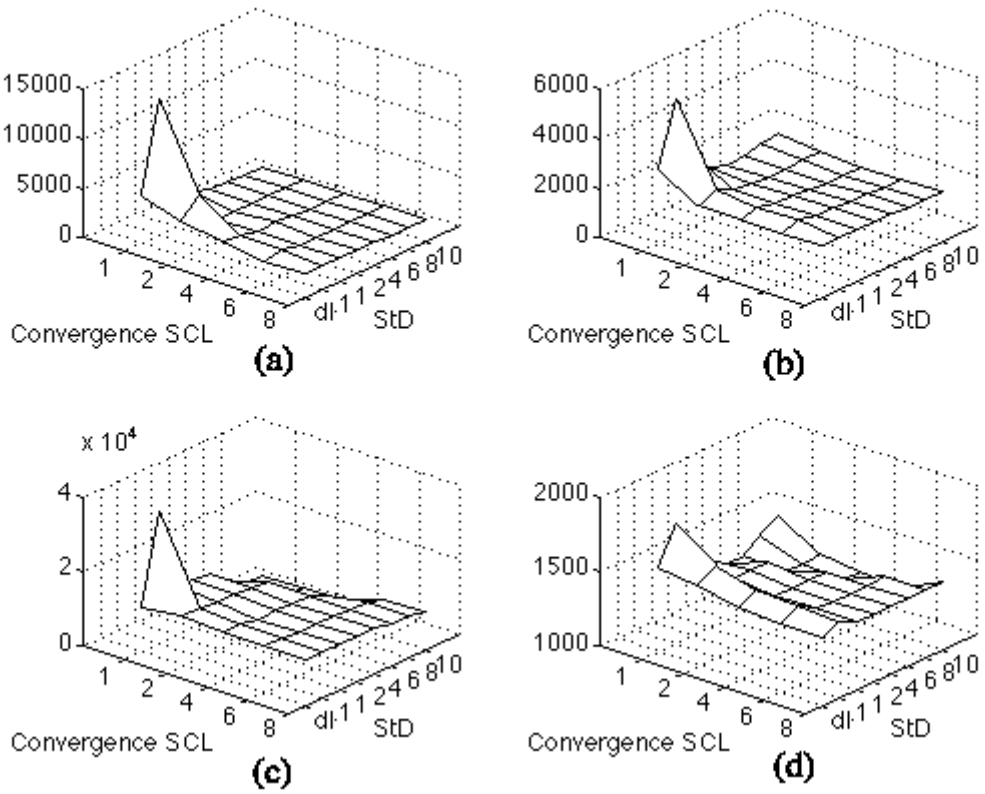


Figura 3.4: Exploración de la sensibilidad del SCS a la desviación estandar inicial ($\sigma_0 = 0.01, 0.1, 1, 2, 4, 6, 8, 10$) y a la velocidad de convergencia al SCL ($r = 1, 2, 4, 6, 8$) . Resultados obtenidos sobre la muestra pequeña comenzando libros de códigos identificados en la figura 3.1

Tabla 3.4: Los vecindarios iniciales del SOM, FLVQ y SCS que producen los resultados en la tabla 3.3.

neighborhood	Sample 1			Sample 2		
	v_0	m_0	σ_0	v_0	m_0	σ_0
Codebook1	3	2	8	2	7	6
Codebook2	1	2	8	1	2	1
Codebook3	3	2	10	1	2	10
Codebook4	2	2	10	1	4	8

Tabla 3.5: Velocidades de convergencia a SCL para SOM, FLVQ y SCS que producen los resultados dados en la tabla 3.3.

r	Sample 1			Sample 2		
	SOM	FLVQ	SCS	SOM	FLVQ	SCS
Codebook1	6	4	1	4	2	8
Codebook2	4	6	4	4	8	6
Codebook3	4	2	1	4	1	1
Codebook4	4	1	6	4	1	4

tamaño del vecindario. En general se puede apreciar que convergencias rápidas a SCL, del orden de $r = 4$, producen buenos resultados para todos los algoritmos. La sensibilidad a los valores iniciales del parámetro de control del vecindario es mayor para los procesos con convergencia lenta a SCL que para los rápidos. Cuando consideramos la convergencia más lenta posible $r = 1$, el SOM es el algoritmo más sensible, produciendo resultados sistemáticamente peores. Por otra parte, FLVQ y SCS son menos sensibles a la velocidad de convergencia a SCL excepto para algunos valores específicos de los respectivos parámetros de control del vecindario. En lo que atañe a estos parámetros, el ajuste óptimo del parámetro v_0 del SOM parece estar relacionado con el tamaño del libro de códigos. El parámetro m_0 del FLVQ está menos relacionado a los datos o tamaños del libro de códigos. Finalmente, la desviación estandard σ_0 del SCS parece estar relacionada con el tamaño de la región ocupada por la muestra de los datos. En términos generales los algoritmos son más sensibles a la velocidad de convergencia a SCL que a los

valores iniciales de los parámetros de vecindad. Entendemos que esto apoya nuestra proposición de utilizar estos algoritmos como si se trataran de procedimientos robustos de inicialización del SCL.

Los mejores resultados se muestran en la tabla 3.3. Los ajustes de los valores iniciales de los parámetros de vecindad y de convergencia a SCL usados para obtener estos resultados están en las tablas 3.4 y 3.5, respectivamente. Los números en negrita corresponden a los resultados que mejoran los obtenidos aplicando el algoritmo de K-medias presentados en la tabla 3.1. Observese que se han obtenido con el SOM. Considerado en conjunto, es el SOM el algoritmo que da los mejores resultados. Las excepciones son, comenzando con el mejor libro de códigos inicial, los resultados de SCS sobre la muestra pequeña y los resultados de SCS y FLVQ sobre la muestra grande. Podemos decir que SOM muestra la más fuerte robustez contra las condiciones iniciales. Sin embargo, el propósito de este experimento no era decidir el algoritmo vencedor entre SOM, FLVQ y SCS, sino mostrar que todos ellos mejoran significativamente sobre SCL, lo que es cierto en la mayor parte de los casos, sobre todo para las condiciones iniciales malas y conforme crece el tamaño de la muestra. Esta mejora será más significativa en el caso no estacionario.

3.3.Resultados experimentales en Cuantización del Color no estacionaria.

La Cuantización del Color de imágenes digitales se puede plantear como un problema de agrupamiento en el espacio 3D definido por el cubo unitario de representaciones RGB [2]. La cuantización del color tiene aplicaciones en visualización y compresión de las imágenes de color [26], [61], [69], [144], [149], [148], segmentación de la imágenes basada en características de color [89], [107], [137] y en la recuperación de imágenes en bases de datos basada en el contenido [71]. Cuando se consideran secuencias de imágenes [40], el planteamiento del problema de la cuantización no estacionaria del color de las secuencias de imágenes asume que la distribuciónm de los colores en las imágenes puede cambiar de forma impredecible. La cuantización no estacionaria del color es, por tanto, un caso de agrupamiento no estacionario en el espacio RGB. Hemos escogido este problema para estudiar la convergencia en un solo paso sobre la muestra por las siguientes

[2]Asumimos que la pérdida perceptiva incurrida por la realización de la cuantización en el espacio RGB es despreciable.

razones:

1. Es un problema técnico realista de dimensiones razonables que permiten la experimentación exhaustiva.

2. Existe un algoritmo casi óptimo, el de mínima varianza de Heckbert, que puede ser utilizado como patrón de referencia para validadr los resultados de los algoritmos basados en redes neuronales competitivas.

En esta sección primero presentamos los datos no estacionario. SObre estos datos hemos aplicado dos algoritmos de referencia: el de mínima varianza de Heckbert y el K-medias (Isodata). El primero da los resultados de referencia óptimos. Tras ello, aplicamos el algoritmo SCL para mostrar la adaptación neuronal básica. Entonces aplicamos el SOM, FLVQ y SCS como modificaciones robustas del SCL frente a condiciones iniciales adversas. Exploramos su robustez contra condiciones iniciales en la secuencia. Finalmente, presentamos algunos experimentos complementarios, dirigidos a clarificar más todavía nuestro punto de vista. Los resultados del algoritmo FLVQ por lotes se comparan con los del FLVQ en linea. Los efectos de considerar vecindarios constantes contra nuestros vecindarios decrecientes exponencialmente se discute en el último experimento.

3.3.1.Los datos experimentales no estacionarios

La secuencia de imágenes usada para el experimento es un *panning* del laboratorio. Las imágenes originales usadas para el experimento tienen una resolución de 480x640 píxeles. Para obtener un cambio suave de la distribución de los colores en las imágenes sucesivas, cada dos imágenes consecutivas se solapan en un 50% de la escena. La distribución de los píxeles en el cubo RGB para las imágenes en la secuencia se muestra en la figura 3.4. Los experimento recojen los resultados de la cuantización a 16 y 256 colores, asumimos que 16 es un valor representativo para tareas de segmentación y que 256 es un número de colores adecuado para tareas de compresión y visualización. Hemos usado los datos de dos maneras. Para el algoritmo de Heckbert, que consideramos el algoritmo de referencia óptimo, los cuantizadores de color se calculan utilizando las imágenes completas. Los restantes algoritmos (Isodata, SCL, SOM, FLVQ, SCS) se aplicaron a muestras de píxeles de las imágenes para calcular los representantes de color, que luego se usan para cuantizar las imágenes completas, y sus resultados de distorsión sobre la imagen completa se comparan con los de Heckbert. De trabajos previos hemos detectado

una cierta sensibilidad de los algoritmos al tamaño de la muestra. Reproduciendo estos experimentos de sensibilidad para todos los algoritmos sería un trabajo ímprobo, por lo que hemos seleccionado tamaños de las muestras adecuados para los tamaños delas tareas: 1600 píxeles para $c = 16$ y 25600 para $c = 256$. Las figuras que muestran los resultados consisten en plots de distorsión a lo largo de la secuencia de imágenes. Todas estas distorsiónes se refieren a la cuantización de las imágenes a tamaño completo. En las cabeceras de los gráficos se muestran la distorsión global, calculada como la suma de las distorsiónes de las imágenes individuales, se da para que se pueda realizar una comparación global de los algoritmos sobre toda la secuencia. Hemos reunido las distorsiónes más significativas sobre las imágenes de tamaño completo en la tabla 3.8. Las magnitudes de las distorsiónes son, obviamente, mayores para la cuantización a 16 colores que para la cuantización a 256 colores.

3.3.2.Los algoritmos de referencia: Heckbert e Isodata

Como algoritmos no adaptativos de referencia hemos usado la versión de mínima varianza del algoritmo propuesto por Heckbert [61] con las mejoras propuestas por [150] para calcular las varianzas locales en cada partición. (De hecho hemos utilizado la implementación que proporciona MATLAB en su *image toolbox*). Este algoritmo realiza las particiones sucesivas del cubo unidad basados en la minimización de la varianza de las particiones resultantes. Implica el cálculo de las varianzas residuales producidas por cada plano de corte, y su complejidad es por tanto del orden del número de particiones producida por la discretización del cubo RGB. Es casi óptimo, pero su costo computacional es muy alto y no puede ser aplicado a problemas de más alta dimensión puesto que su complejidad crece exponencialmente con la dimensionalidad del espacio de los datos. Por otra parte, nuestros algoritmos adaptativos neuronales tienen una complejidad que crece linealmente con el número de representates de color, el tamaño de la muestra y la dimensión del espacio de los datos. Esto significa que nuestra aproximación se puede extender a problemas de más alta dimensión mientras que el algoritmo de Heckbert no puede serlo.

Las figuras 3.4a,b muestran los resultados de la aplicación del algoritmo de Heckbert a los datos no estacionarios, buscando el cuantizador óptimo a 16 y 256 colores. Este algoritmo se ha aplicado a las imágenes completas de la secuencia de dos maneras. En primer lugar, asumiendo la naturaleza no estacionaria de los datos se ha aplicado a cada imagen independientemente, lo que nos da los mejores

resultados en las figuras 3.4a,b. La curva correspondiente se denota Time Varying en las figuras. En segundo lugar se considera que los datos son estacionarios, por lo que en principio un libro de códigos calculado para la primera imagen daría buenos resultados al cuantizar las restantes imágenes en la secuencia. Los resultados de aplicar esta estrategia se denotan como Time Invariant en las figuras 3.4a,b. Las distorsiónes globales en la secuencia (vease la tabla 3.8) son mucho mayores que en las aplicaciones variantes en el tiempo del algoritmo.

Las aplicaciones *Time Varying* y *Time Invariant* del algoritmo de Heckbert son útiles para definir las cotas del comportamiento adaptivo en los experimentos. Nos dan las cotas superiores e inferiores para los restantes algoritmos. Un algoritmo no se puede considerar adaptativo si su curva de distorsión a lo largo de la secuencia es mayor en algún punto que la curva correspondiente a la realización *Time Invariant* del algoritmo de Heckbert. Por otra parte, la curva correspondiente a la realización *Time Varying* del algoritmo de Heckbert es la mejor respuesta que esperamos de cualquier algoritmo adaptativo.

En las figuras 3.4c y 3.4d mostramos los resultados de la aplicación del algoritmo k-medias como un algoritmo adaptativo de la forma que se discute en la sección 2.7.4. El libro de códigos inicial para la secuencia de imágenes es el obtenido mediante el algoritmo de Heckbert para la primera imagen y la adaptación comienza en la segunda imagen. El criterio de parada del algoritmo de k-medias es que la diferencia absoluta entre las distorsiónes en interaciones consecutivas sea menor que 0.01. Como la curva de los resultados del Isodata cae dentro de la región definida por los resultados de la aplicación invariante y variante en el tiempo del algoritmo de Heckbert, podemos decir que se comporta como un algoritmo adaptativo puesto que su curva es claramente mejor que la corresondiente a la aplicación invariante en el tiempo (Time Invariant) del algoritmo de Heckbert. En algunos instantes la cuantización a 16 colores con el Isodata alcanza los resultados de referencia de la aplicación variante en el tiempo del algoritmo de Heckbert, que es nuestro valor de referencia. Las distorsiónes globales mostradas en la tabla 3.8 son cercanas a la referencia óptima. El número de iteraciones necesarios para cada imagen en la secuencia se muestra en la figura 3.4f. Se puede aprecialr que para las imágenes que muestran los mayores cambios en la distribución de los colores, el Isodata necesita un número mayor de iteraciones para alcanzar el criterio de parada. El número de iteraciones necesitadaspor cada imagen en la secuencia se muestra en las figuras 3.4e y 3.4f, para 16 y 256 colores respectivamente. Se puede apreciar que para las imágenes que muestran mayores cambios en la distribución del color, el algoritmo Isodata necesita un número mayor de iteraciones

para alcanzar la condición de parada. También, el incremento en el número de agrupamientos y el tamaño de la muestra de los píxeles de la imagen implica el aumento en el número de interaciones que necesita Isodata para alcanzar la condición de parada. Dado que el entrenamiento de las redes competitivas será en un único paso sobre la muestra, hemos incluido también en la figuras 3.4g y 3.4h los resultados de un único ciclo del Isodata. Naturalmente la degradación de los resultados es mayor para las imágenes con mayor variación de la distribución de color. La degradación global introducida por la aplicación en un único paso se puede apreciar en la tabla 3.8.

La aplicación del Isodata muestra una sensibilidad al número de colores buscados que es común a todos los algoritmos que hemos probado. Los resultados son cualitativamente peores para el caso de 256 colores que para el caso de 16 colores. La calidad de los algoritmos se mide por la distancia de la respuesta del algoritmo a relativa a la respuesta óptima (Time Varying). Es evidente en figuras 3.4c,d que esta distancia relativa aumenta al aumentar el número de agrupamientos buscados.

Tabla 3.6: Exploración de la sensibilidad en el caso de 16 colores. Distorsión global de la cuantizacion de las muestras de tamaño $n = 1600$ de cada imagen de la secuencia tras el calculo de los libros de códigos con SOM, FLVQ y SCS bajo las diferentes combinaciones de ajustes de los vecindarios iniciales y la velocidad de convergencia a SCL

	SOM $\upsilon_0 =$		$FLVQ$ $m_0 =$				SCS $\sigma_0 =$		
	1	8	10	7	4	2	0.1	2	$\widehat{\sigma_{i,0}}$
$r = 1$	102.2	106.2	99.07	96.87	93	84.04	95.31	442.2	236.2
$r = 2$	72.08	71.49	91.47	90.94	86.96	84.74	85.85	149.2	85.24
$r = 4$	70.8	70.15	85.34	85.91	84.87	84.56	81.77	125.8	85.62
$r = 6$	72.7	69.37	85.39	85.31	85.48	84.66	81.67	113.7	84.43
$r = 8$	74.11	70.48	87.18	86.15	87.22	86.01	82.43	108.8	82.57
$fixed\ \Phi$	102.1	774.3	382.8	382.8	351.1	94.57	256.3	3819	187.4

Tabla 3.7: Exploración de la sensibilidad en el caso de 256 colores. Distorsión global de la cuantizacion de las muestras de tamaño $n = 25600$ de cada imagen de la secuencia tras el calculo de los libros de códigos con SOM, FLVQ y SCS bajo las diferentes combinaciones de ajustes de los vecindarios iniciales y la velocidad de convergencia a SCL

	SOM $v_0 =$		$FLVQ$ $m_0 =$		SCS $\sigma_0 =$	
	8	128	7	2	0.1	$\widehat{\sigma_{i,0}}$
$r = 1$	385.5	394.8	387.5	350.5	439.5	527.5
$r = 2$	302.4	304.5	351.8	341.2	361.7	359.6
$r = 4$	294.2	294.6	352.2	346.3	341.9	352
$r = 6$	295.5	291.2	343.4	349	329.4	349.8
$r = 8$	299.2	288.7	357	349.9	323.1	355.6
$fixed\ \Phi$	763.2	13650	784.9	465.1	2653	561.3

3.3.3.El aprendizaje competitivo simple (SCL).

En la figura 3.4 mostramos los resultados de la aplicación de la adaptación en un sólo paso sobre la muestra con SCL, con las programaciones de la velocidad de aprendizaje discutida arriba, empotrada en los resultados del algoritmo de Heckbert. En las figuras 3.4a,b el libro de códigos inicial es el obtenido por el algoritmo de Heckbert en la primera imagen y la adaptación comienza en la segunda imagen. Se puede apreciar que SCL se comporta como un algoritmo adaptativo en el sentido de que mejora a la aplicación invariante en el tiempo del algoritmo de Heckbert, pero es menos óptimo que la aplicación variante en el tiempo del mismo algoritmo. Se puede apreciar también que mejora sobre la aplicación en un sólo paso del algoritm Isodata (vease la tabla 3.8 para comparar las distorsiónes globales en ambos casos). Para resaltar la distancia relativa al óptimo mostramos en las figuras 3.4c,d la distorsión relativa de SCL respecto de las curvas obtenidas con las aplicaciones variante e invariante en el tiempo, respectivamente, del algoritmo de Heckbert. Estas curvas se calculan como sigue

$$SCL_{relative}\,(\#i) = \frac{SCL\,(\#i) - \ \mathsf{Time\ Varying}\,(\#i)}{\mathsf{Time\ Invariant}\,(\#i) - \mathsf{Time\ Varying}\,(\#i)}; i = 2, 3, ... \qquad (3.9)$$

Por tanto $SCL_{relative}\,(\#i)$ es negativa cuando SCL mejora a la aplicación variante en el tiempo del algoritmo de Heckbert (nuestro casi-óptimo de referencia). Esta cantidad es mayor que 1 cuando la respuesta de SCL es peor que la aplicación invariante en el tiempo del algoritmo de Heckbert. Está claro por la comparación de 3.4c y 3.4d que el aumento en el tamaño del libro de códigos degrada los resultados de SCL también.

La condición inicial para la secuencia, el libro de códigos de Heckbert para la primera imagen, es bastante buena, en las figuras 3.4e,f probamos la respuesta a otras condiciones iniciales. Estas se asumen como los libros de códigos para la primera imagen en la secuencia y se usan para comenzar el proceso de adaptación en las siguientes imágenes. Los distintos libros de códigos iniciales probados son:

- in Sample que es un buen libro de códigos inicial extraido de la muestra de la primera imagen. No coinciden en los casos de 16 y 256 colores.
- in RGB box es un libro de códigos arbitrario generado aleatoriament en el cubo RGB. Es la peor condición inicial.
- Umbral corresponde a una selección guiada por una condición de umbral entre los elementos de la muestra de la primera imagen.

Se puede apreciar que para las buenas condiciones iniciales, SCL se recupera tras la segunda imagen. Sin embargo, permanece rindiendo peor que cuando comienza por el libro de códigos calculado por el algoritmo de Heckbert (figuras .3.4a,b). Las peores respuestas corresponden a las peores condiciones iniciales en las que el efecto de la mala inicialización se propaga a través de toda la secuencia de imágenes. Mostraremos al reproducir estos experimentos más adelante que las redes competitivas consideradas mejoran la robustez frente a condiciones iniciales del SCL.

3.3.4.Sensibilidad a los parámetros de control del SOM, FLVQ y SCS

Comenzamos el estudio experimental de SOM, FLVQ y SCS realizando un experimento de sensibilidad similar al discutido para el caso estacionario. La condición inicial es el libro de códigos de Heckbert en la primera iamgen, y la adaptación comienza en la segunda imagen como se propone en la sección 2.7.4. El experimento de sensibilidad se restringe a los resultados de distorsión sobre la muestra. A partir de estos resultados decidimos los valores óptimos de los parámetros de control de las funciones de vecindad en cada caso. Usamo el libro de códigos

correspondiente para realizar al cuantización de las imágenes a tamaño completo, asumiendo que dicha optimalidad se extrapolará de la muestra a la imagen de tamaño completo.

La tabla 3.6 muestra los resultados de la distorsión global en la secuencia de muestras de 1600 píxeles cuantizadas a 16 colores (observese que la magnitud de la distorsión global es menor que en la tabla 3.6). La última fila en la tabla proporciona los resultados de distorsión obtenidos con parámetros de vecindad constantes. Los malos resultados son debidos al hecho de que los mínimos locales encontrados por las redes neuronales lo son de sus respectivas funciones objetivo para un valor fijo de los parámetros de vecindad. Estas funciones son bastante distintas de la distorsión euclídea que minimiza el algoritmo de Heckbert, el Isodata y el SCL. La inspección del resto de la tabla muestra que los peores resultados se obtienen cuando no hay una fase SCL en el aprendizaje ($r = 1$). En general, la convergencia rápida a SCL mejora los resultados. En el SOM el radio inicial es un factor secundario del rendimiento. En el FLVQ, el exponente inicial es mas significativo cuando $m_0 = 2$, para $m_0 > 2$ el FLVQ es bastante insensible a este parámetro. En el SCS el efecto de la desviación estandard inicial es muy fuerte. Las estimaciónes locales $\widehat{\sigma}_{i,0}$ dan buenos rendimiento, aunque no los óptimos. Las velocidad de convergencia óptima a SCL es $r = 6$ en todos los casos. Los parámetros de vecindario iniciales óptimos son $v_0 = 8$, $m_0 = 2$ y $\sigma_0 = 0.1$.

La tabla 3.7 muestra la distorsión global resultante de la secuencia de muestras de 25.600 píxeles cuantizadas a 256 colores. Otra vez la última fila da los resultados para vecindarios fijos que confirman la necesidad de realizar al convergencia a SCL para que produzca la minimización de la distorsión, aunque el efecto es menos notorio para FLVQ. El rendimiento de las restantes entradas en la tabla mejora con el incremento en la velocidad de convergencia a SCL. Los valores iniciales de la función de vecindad son factores secundarios en el rendimiento. Las velocidades de convergencia a SCL óptimas son $r = 8, 2, 8$ para el SOM, FLVQ y SCS, respectivamente. Los valores iniciales óptimos son $v_0 = 128$, $m_0 = 2$ y $\sigma_0 = 0.1$.

La figura 3.4 muestra los resultados de la aplicación de SOM, FLVQ y SCS bajo los ajuste óptimos de los parámetros de control de cada algoritmo deducidos de las tablas 3.6 y 3.7. Los resultados de sensibilidad previos tienen el propósito de garantizar que cada algoritmo se aplica en sus mejores ajustes particulares. Los resultados de distorsión global se pueden encontrar en la tabla 3.8 correspondiendo a la inicialización de Heckbert.

Las figuras 3.4a,b muestran la distorsión de la cuantización a 16 y 256 colores

de cada imagen a tamaño completo en la secuencia de imágenes. EL SOM da los mejores resultados. Esta afirmación es más clara si consideramos la distorsión global: el SOM en un paso sobre la muestra mejora incluso al Isodata con muchas iteraciones. En general, los tres algoritmos mejoran sobre SCL y el algoritmo Isodata en una iteración. Las figuras 3.4c,d muestran la distorsión relativa (calculada de forma similar a la ecuación 3.9). Se puede apreciar que el SOM en un paso sobre la muestra encuentra a veces mejores representantes de color que la aplicación variante en el tiempo del algoritmo de Heckbert. Finalmente, para resaltar la mejora sobre SCL, las figuras 3.4e,f muestran la substracción de las curvas de cada algoritmo de la correspondiente a SCL. Los tres algoritmos muestran mejoras significativas.

3.3.5.Sensibilidad a las condiciones iniciales

Estos experimentos son una extensión de los presentados en [49] y [50]. En las secciones anteriores, el proceso adaptativo comienza en la segunda imagen, asumiendo como libro de códigos inicial el obtenido por el algoritmo de Heckbert en la primera imagen, lo que representa una condición inicial muy buena. En este apartado consideramos la respuesta de los algoritmos SOM, FLVQ y SCS como AVQ ante condiciones iniciales peores y los comparamos con los resultados obtenidos con SCL en el apartado 3.3.3. Las condiciones iniciales son las mismas que las utilizadas para calcular las figuras 3.4e,f del apartado 3.3.3. El ajuste de las funcione de vecindad es el mismo que el aplicado para obtener la figura 3.4. Como en esas figuras los resultados ploteados son la distorsión de la cuantización de color cada imagen a tamaño completo en la secuencia con los libros de códigos obtenidos sobre las muestras de las imágenes. La distorsión global se reproduce en la tabla 3.8.

La figura 3.4a,b da los resultados del FLVQ con 16 y 256 representantes de color, respectivamente. Se puede apreciar que las mejoras respecto del SCL son menores. La figura 3.4c,d da los resultados del SCS con 16 y 256 representantes de color, respectivamente. Se puede apreciar una mejora sistemática respecto del SCL, aunque no muy grande. Finalmente, la figura 3.4e,f da los resultados del SOM con 16 y 256 representantes de color, respectivamente. Se puede apreciar su sorprendente robustez frente a condiciones iniciales y los resultados son claramente los mejores. Mientras que para FLVQ y SCS los efectos de las condiciones iniciales se propagan a lo largo de toda la secuencia, el SOM se colapsa casi completamente al comportamiento óptimo desde la segunda imagen en la secuencia. Esta robustez

tiene implicaciones prácticas para procesos de video en tiempo real puesto que permite inicializaciones arbitrarias.

3.3.6.FLVQ online versus batch

La proposición original del FLVQ presentada en [8] es un algoritmo batch en el que el exponente se actualiza tras cada iteración de cálculo de los representantes borrosos de los agrupamientos por las ecuaciones (2.57) y (2.58). El exponente decrece linealmente aproximandose a 1 por arriba (en la práctica 1.1 se considera equivalente a 1). Hasta este momento nuestro trabajo con FLVQ se adhiere a lo propuesto por los autores. Sin embargo se puede argüir que la aplicación online en un solo paso propuesta aquí degrada el rendimiento del algoritmo. De hecho, ha habido un debate duro sobre la definición apropiada del algoritmo FLVQ online [108] [18] [111].

Queremos verificar la degradación en rendimiento introducida por nuestra definición online de FLVQ. Para esto, comparamos los resultados de la realización online con los resultados de la realización batch. En la primera hemos aplicado los ajustes de los parámetros de control sugeridos en nuestro experimento de sensibilidad. En la segunda hemos aplicado la programación general del exponente sugerida en [8]. La figura 3.4a,b muestra el rendimiento en el caso de 16 colores para $m_0 = 7$ y $m_0 = 2$, respectivamente. El libro de códigos inicial fue el obtenido por el algoritmo de Heckbert sobre la primera imagen. La adaptación se realizó sobre la muestra de 1600 píxeles, y el resultado mostrado es la distorsión por cada imagen de la cuantización de las imágenes a tamaño completo. Los resultados se presentan normalizados en relación a los del algoritmo de Heckbert para resaltar las diferencias.

Hemos encontrado que la versión batch de FLVQ mejora la versión online. Sin embargo, el costo computacional de FLVQ batch es tan alto que hemos sido incapaces de realizar el experimento sobre 256 colores según la planificación que nos habíamos propuesto.

3.3.7.Vecindarios constantes

Hemos estado enfatizando la relación entre las funciones de vecindad y la función objetivo que está siendo efectivamente minimizada por el algoritmo de redes neuronales competitivas. Hemos dado en las tablas 3.6 y 3.7 la distorsión global de las cuantizaciones de las muestras resultantes de realizar la adaptación con vecindarios constantes. Para hacer más evidente el efecto del vecindario constante,

presentamos en la figura 3.4 la distorsión en cada imagen a tamaño completo. Comparamos los mejores resultados de SOM, FLVQ y SCS en la figura 3.4 con los obtenidos congelando el vecindario en su tamaño inicial óptimo. Denotamos SOMFIX, FLVQFIX y SCSFIX las curvas de estos resultados en los gráficos.

La degradación observada es mayor para 256 representantes de color que para 16. El SOM se convierte en un algoritmo no adaptativo en el caso de 256 representantes de color con vecindario fijo. El FLVQ parece ser el menos afectado por la congelación del vecindario. La razón es que el criterio borroso es muy cercano a la distorsión euclídea cuando $m = 2$. El algoritmo SCS se degrada dramaticamente en el caso de 16 representantes de color. Los algoritmos se comportan de acuerdo con sus funciones objetivo, cuando son cercanas a la distorsión euclídea sus resultados no se degradan mucho, como en el caso de FVLQ. La función objetivo inducida por la constante $v = 128$ para el SOM es muy diferente de la distorsión euclídea y consecuentemente los resultados son muy malos, aunque el SOM puede estar efectivamente minimizando su función objetivo. Aunque pueden parecer triviales, estos ejemplos ilustran los efectos producidos por muchos algoritmos heurísticos de agrupamiento y cuantización vectorial y la dificultad de realizar una comparación precisa de sus rendimeintos ya que de hecho están realizando la minimización de funciones objetivo distintas.

3.4. Conclusiones

En este capítulo hemos propuesto una programación novedosa de los parámetros de aprendizaje de tres redes neuronales competitivas: SOM; FLVQ y SCS. Esta programación permite su aplicación eficiente como mecanismo de cuantización vectorial adaptativa para la solución de tareas de Agrupamiento no estacionario. Como tarea experimental hemos escogido la cuantización del color de secuencias de imágenes.

Como resultado de los experimentos computacionales realizados en este capítulo, el SOM parece ser el algoritmo más robusto y eficiente. Los experimentos de sensibilidad también muestran que estrategias muy generales pueden dar lugar a ajustes eficientes de los parámetros de aprendizaje. En general una tasa de convergencia a SCL $r = 4$ da buenos resultados para todos los algoritmos en los experimentos presentados aquí y en otros lugares. Para el SOM, un radio de vecindario inicial $v = \frac{c}{2}$ es una buena regla. Para el FLVQ un exponente inicial $m_0 = 2$ da buenos resultados. Finalmente, el SCS es muy sensible al tamaño de la región del espacio ocupado por los datos. Una buena estrategia consiste en calcular la desviación

estandard a partir de las distancias entre los vectores código, como un las bolas sin solapamiento de 95% de confianza en torno a ellos.

Dentro de los planes de trabajo futuro, planeamos trabajar en la extensión de la aproximación presentada aquí hacia otros algoritmos neuronales y borrosos, como la familia FALVQ propuestta por [72]. Otra línea de trabajo es la aplicación de nuestra programación a problemas de más alta dimension, y empotrar la cuantización de color no estacionaria en sistemas reales de proceso de video. Nuestra aproximación es computacionalmente competitiva: su complejidad crece linealmente con la dimensión del espacio y los tamaños de la muestra de vectores.

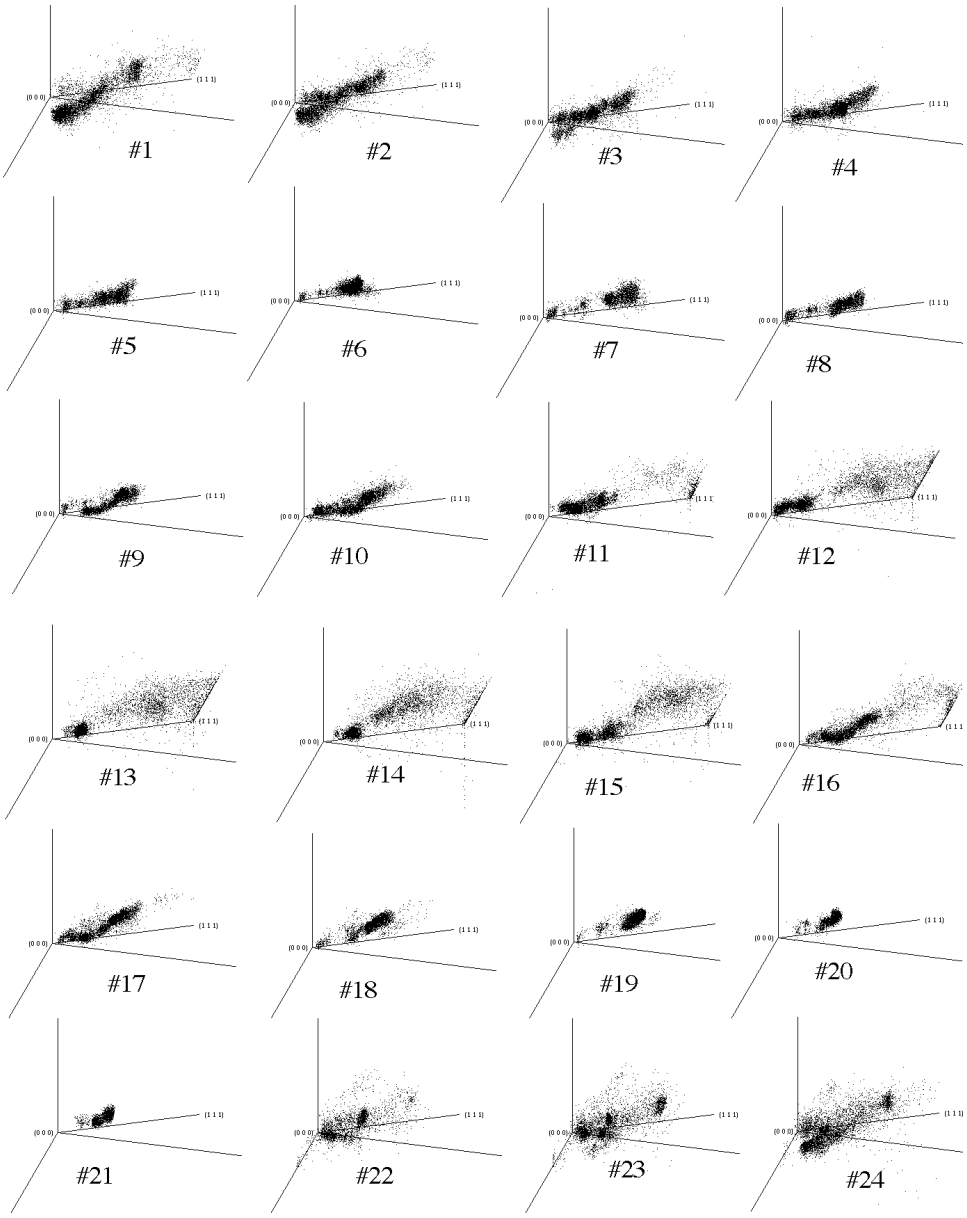


Figura 3.5: Datos no estacionarios. Distribución de los colores de los píxeles en el cubo RGB para cada imagen en la secuencia experimental

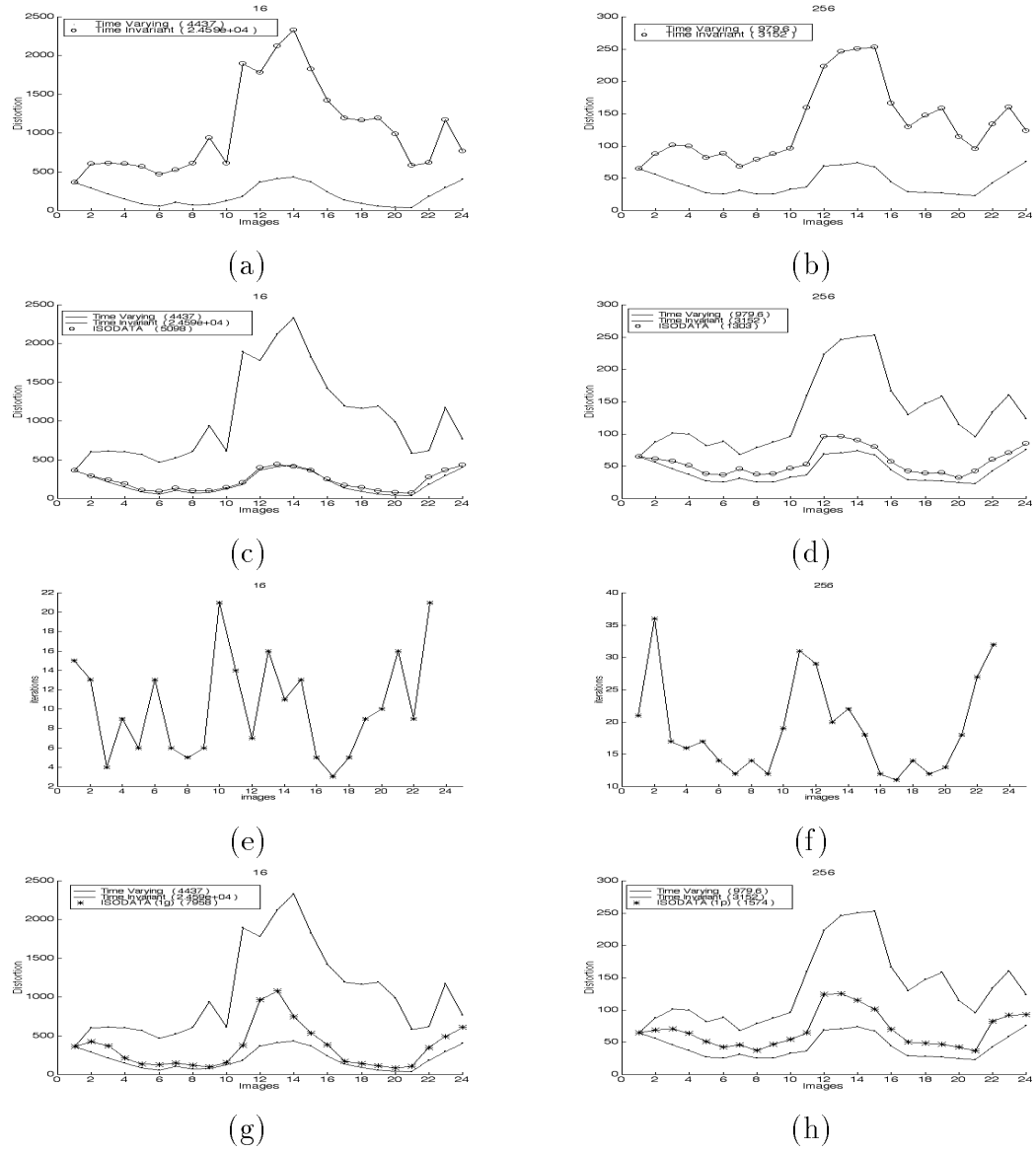


Figura 3.6: Resultados de distorsión para los libros de códigos calculados con los algoritmos de referencia. La cuantizacion a 16 colores en (a,c,e,g) y a 256 colores en (b,d,f,h). (a,b) Los resultados de referencia obtenidos con el algoritmo de Heckbert. (c,d) Los resultados obtenidos con el algoritmo de k-medias con el criterio de parada dado en el texto. (e,f) el Número de iteraciones necesitados por el k-medias para alcanzar el criterio de parada. (g,h) resultados de distorsión de los libros de códigos obtenidos con una iteracion del algoritmo k-medias sobre cada imagen

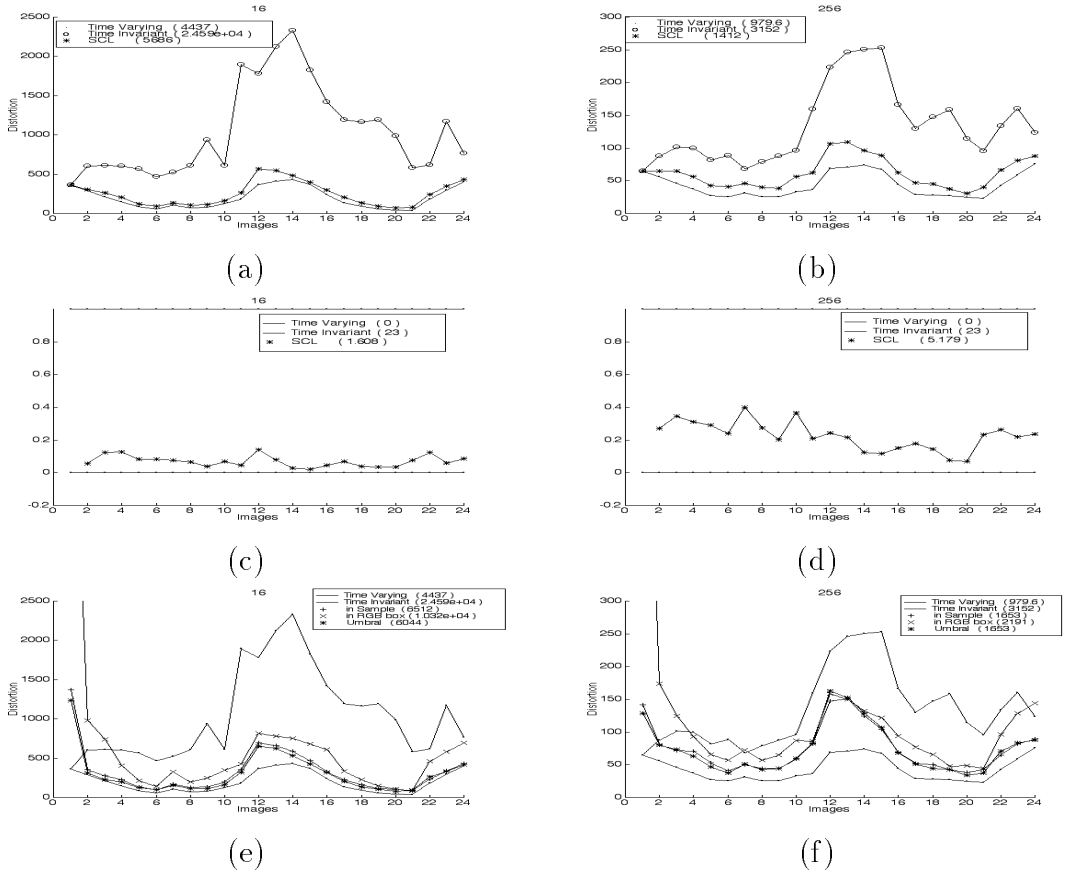


Figura 3.7: Distorsión en cada imagen resultante de la cuantizacion de las Imágenes a tamaño completo con los libros de códigos calculados por el SCL en las muestras de las Imágenes. (a,c,e) 16 representantes de color y muestras de 1600 píxeles. (b,d,f) 256 representantes de color y muestras de 25600 píxeles. (a,b) Resultados de distorsión. (c,d) Resultados de distorsión relativos. (e,f) resultados de sensibilidad comenzando por libros de códigos diferentes del proporcionado por el algoritmo de Heckbert en la primera imagen.

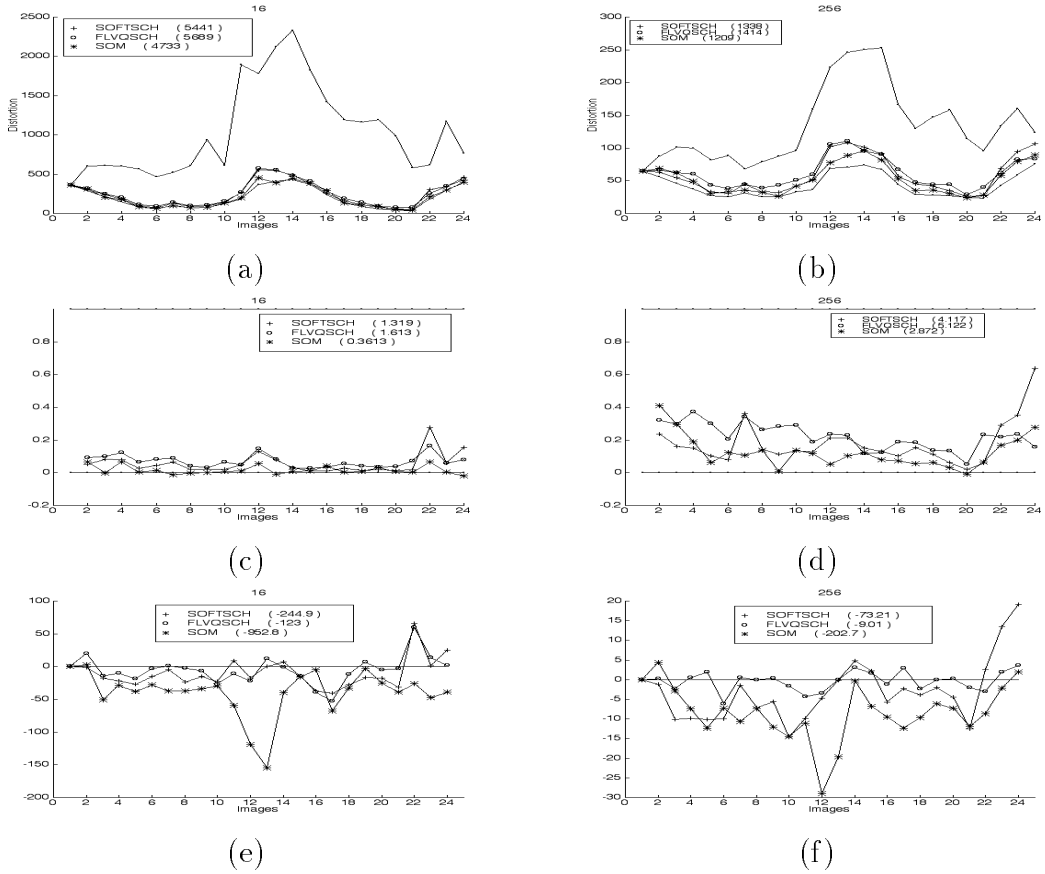


Figura 3.8: Distorsión en cada imagen resultante de la cuantizacion de las Imágenes a tamaño completo con los libros de códigos calculados por el SOM, FLVQ y SCS con ajustes optimos de los parámetros de vecindad deducidos de las tablas 3.6 y 3.7. (a,c,e) 16 representantes de color y muestras de 1600 píxeles. (b,d,f) 256 representantes de color y muestras de 25600 píxeles. (a,b) Resultados de distorsión. (c,d) Resultados de distorsión relativos. (e,f) substracion en cada imagen de la distorsión obtenida por SCL.

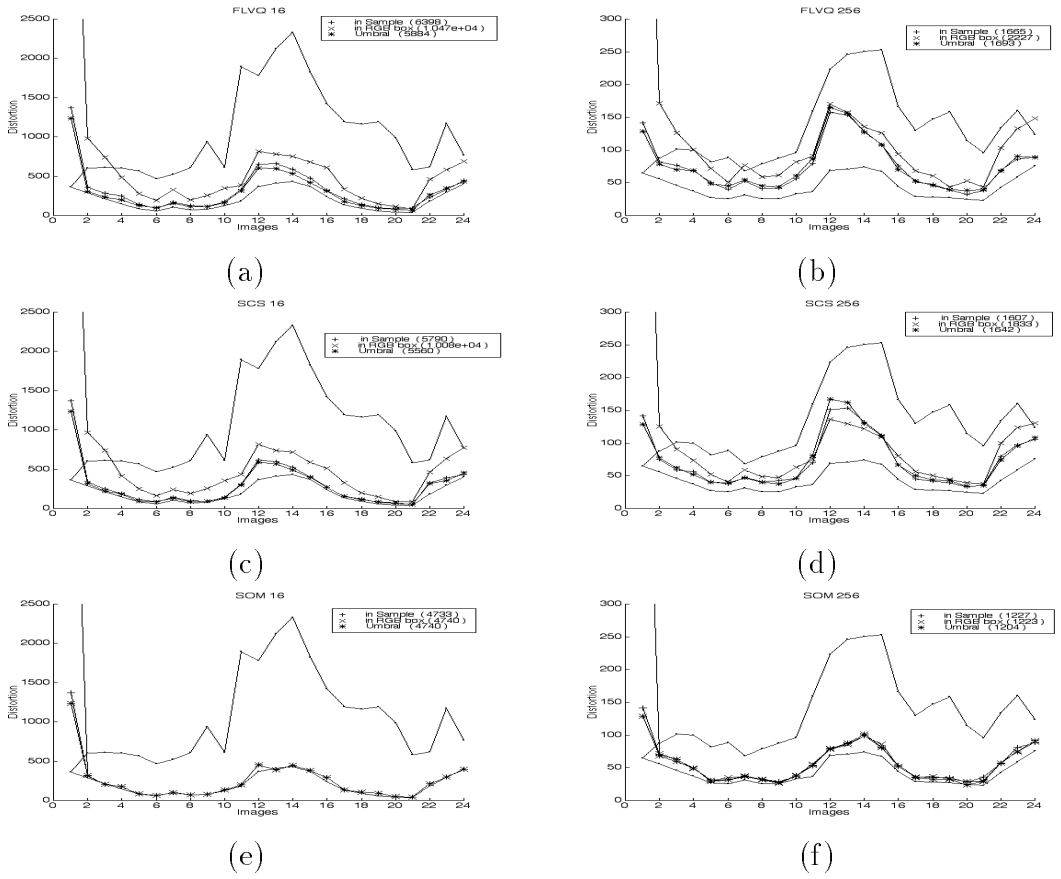
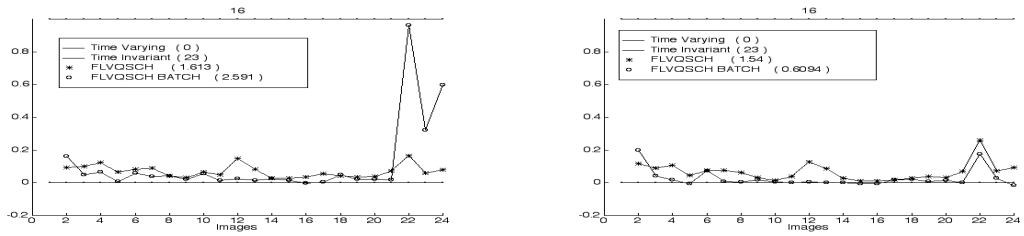


Figura 3.9: Distorsión en cada imagen que muestran la sensibilidad a las condiciones iniciales del SOM, FLVQ y SCS. Los libros iniciales de códigos se escogen como se indica en el texto. Los parámetros de vecindad se ajustan como en la figura 3.8 (a,c,e) 16 representantes de color y muestras de 1600 píxeles. (b,d,f) 256 representantes de color y muestras de 25600 píxeles. (a,b) Resultados del FLVQ. (c,d) Resultados del SCS. (e,f) Resultados del SOM.



(a) (b)

Figura 3.10: Aplicaciones online versus aplicaciones batch de FLVQ. Resultados de distorsión relativa por imagen de la cuantizacion a 16 colores calculado sobre la muestra de 1600 píxeles. (a) con exponente inicial $m_0 = 7$ y (b) con $m_0 = 2$

Tabla 3.8: Distorsión global de las secuencias de Imágenes por la cuantizacion de las Imágenes completas usando los libros de códigos calculados sobre las muestras (excepto en el caso del algoritmo de Heckbert). La inicializacion refleja los libros de códigos iniciales usados para la secuencia completa. Heckbert denota el libro de codigo de Heckbert para la primera imagen de la secuencia

Algorithm	Initialization	16	256
Time Invariant		24590	3152
Time Varying		4437	979.6
Isodata	Heckbert	5098	1303
Isodata (1 iter.)	Heckbert	7958	1574
SCL	Heckbert	5686	1412
	in Sample	6512	1653
	in RGB box	10320	2191
	Umbral	6044	1653
SOM	Heckbert	4733	1209
	in Sample	4733	1227
	in RGB box	4740	1223
	Umbral	4740	1204
FVLQ	Heckbert	5689	1414
	in Sample	6398	1665
	in RGB box	10470	2227
	Umbral	5884	1693
SCS	Heckbert	5441	1338
	in Sample	5970	1607
	in RGB box	10080	1833
	Umbral	5560	1642

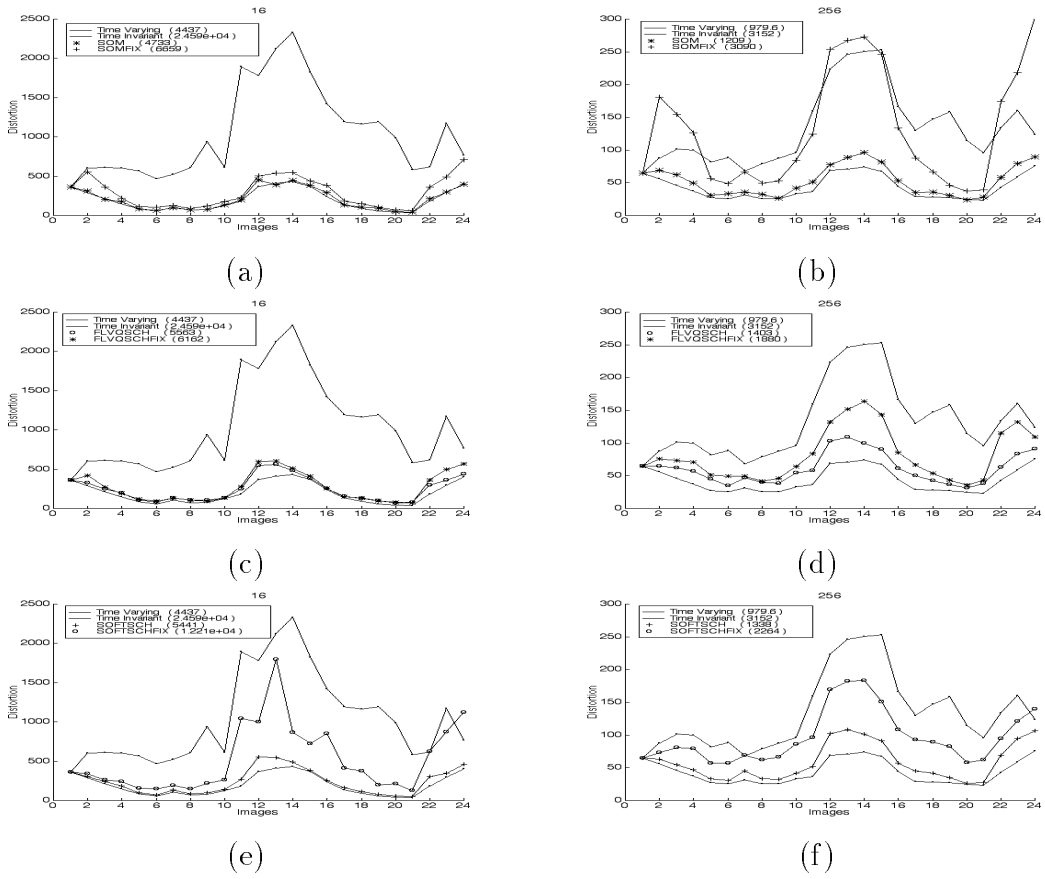


Figura 3.11: El efecto de vecindarios constantes. Distorsión por imagen de la cuantizacion de la imagen completa con los libros de códigos calculados a partir de las muestras. (a,c,e) 16 representantes de color sobre la muestra de 1600 píxeles y (b,d,f) 256 representantes de color sobre las muestras de 25600 píxeles. Los ajustes de los parámetros en la figura 3.8 versus (a) $v = 8$, (b) $v = 128$, (c,d) $m = 2$, (e,f) $\sigma = 0.1$

4. Filtrado basado en la cuantización vectorial

La aportación central de la tesis consiste en la definición de un algoritmo de filtrado y segmentación de la imagen basado en la cuantización vectorial de bloques de la imagen. En este capítulo describimos con detalle esta aproximación en el marco del procesado bayesiano de la imagen, mostrando que puede ser aplicado para recuperar deformaciones locales en los vecindarios de los píxeles. El algoritmo se denomina VQ-BF como abreviatura de Filtrado Bayesiano basado en Cuantización Vectorial.

En la sección 4.1 presentamos la notación utilizada. En la sección 4.2 se presenta el modelo de los bloques de píxeles. En la sección 4.3 se presenta el modelo total del VQ-BF. En la sección 4.4 se hacen algunas consideraciones sobre la preservación de los bordes en las imágenes filtradas con VQ-BF. En las sección 4.5 se presenta el SOM como el algoritmo más apropiado para la realización de la cuantización vectorial. Finalmente, en la sección 4.6 se presentan las conclusiones del capítulo.

4.1.Notación

Comenzaremos recordando la notación estandar en el procesado bayesiano de la imagen [41], [147]. Una imagen puede describirse por una tupla $x = \left(x^P, x^L, x^E, ..\right)$ cuyos componentes corresponden a los atributos de interés para la aplicación: la intensidad x^P, las etiquetas de clasificación de los pí xeles x^L, las posiciones de los bordes x^E y otros. Sea S^P una malla cuadrada finita donde cada sitio representa una posición de un pixel de la imagen. El vector $x^P = \left(x_s^P\right)_{s \in S^P}$ representa un patrón de configuración de los valores de gris en los píxeles. Si tratamos de clasificar los píxeles de la imagen asignandoles un significado preciso, ya sea en aplicaciones de análisis de textura como en otras, $x^L = \left(x_s^L\right)_{s \in S^L}$ es un patrón de etiquetas asociadas a los bloques de píxeles en el conjunto S^L, de forma que

$x_s^L = l \in L$ es la etiqueta de clase del bloque s. Estos bloques pueden superponerse o no.

Los datos observables y son una función Y de la imagen auténtica x. El espacio de los datos observables se denota por $\mathbf{Y}$ y el de las imágenes auténticas por $\mathbf{X}$. Dado $x \in X$, la ley probabilística que modela Y se denota $\mathsf{P}(y|x)$. SI $\mathbf{Y}$ es finita, $\mathsf{P}(y|x)$ es la probabilidad de observar y cuando la imagen auténtica es x. Por tanto, para cada $x \in \mathbf{X}$, $\mathsf{P}(y|x)$ es una distribución de probabilidad sobre $\mathbf{Y}$. i.e. $\mathsf{P}(y|x) \geq 0$ y $\sum_y \mathsf{P}(y|x) = 1$.

Las expectaciones a priori se pueden formular como restricciones sobre la imagen ideal. La función positiva y normalizada $\Pi(x)$ definida en el espacio de las imágenes $\mathbf{X}$ es la distribución *a priori*. La elección del modelo *a priori* de pende del problema y es uno de los pasos clave en el análisis Bayesiano de la imagen. La distribución *a priori* Π y las probabilidades condicionales determinan la distribución conjunta de los datos y de las imágenes en el espacio producto $\mathbf{X} \times \mathbf{Y}$ por

$$\mathsf{P}(x,y) = \Pi(x)\,\mathsf{P}(y|x), x \in \mathbf{X}, y \in \mathbf{Y}. \tag{4.1}$$

Esta es la ley probabilística de un par de variables aleatorias (X,Y) en $\mathbf{X} \times \mathbf{Y}$ donde X tiene una ley Π e Y tiene una ley Γ dada por $\Gamma(Y=y) = \sum_x \mathsf{P}(x,y)$. La probabilidad *a posteriori* de $x \in \mathbf{X}$ está dada por

$$\mathsf{P}(x|y) = \frac{\Pi(x)\,\mathsf{P}(y|x)}{\sum_z \Pi(z)\,\mathsf{P}(y|z)}. \tag{4.2}$$

Para datos continuos, la distribución a priori Π tendrá en general la forma de la distribución de Gibbs

$$\Pi(x) = Z^{-1}\exp(-H(x)), Z = \sum_{z \in \mathbf{X}} \exp(-H(z)). \tag{4.3}$$

Donde H es la función de energía de Π. En la mayor parte de los casos, el modelo *a posteriori* es también una función de Gibbs, i.e. hay una función $H(\cdot|y)$ definida en un subespacio $\widehat{\mathbf{X}}$ de $\mathbf{X}$ tal que

$$\mathsf{P}(x|y) = Z^{-1}(y)^{-1}\exp(-H(x|y)), x \in \widehat{\mathbf{X}}. \tag{4.4}$$

La función de energía está conectada con la función de verosimilitud. La energía *a posteriori* se puede escribir de la siguiente manera:

$$H(x|y) = \widetilde{c}(y) - \ln \mathsf{P}(y|x) + H(x) \tag{4.5}$$

Un modo de la distribución a posteriori $\widehat{x} = \max_{x} \{\mathsf{P}(x\,|y)\}$ es el estimador Máximo a Posteriori (MAP) de x dado y. Si el procesado se realiza independientemente en todos los sitios de la imagen, el modo de la estimación marginal posterior $\widehat{x}_s$ maximiza la distribución posterior marginal $\mathsf{P}(x_s\,|\widehat{y})$. Es bien sabido que los estimadores MAP son los estimadores óptimos Bayesianos para la función de pérdida $0-1$.

La maximización de las distribuciones de Gibbs es equivalente a la minimización de su función de energía, por tanto, el estimador MAP se puede calcular sin evaluar la función de partición Z. Sea la observación de la forma $Y = \varphi(X, \eta)$. La ley probabilistica del ruido η se denota por Γ. Si el ruido η y la imagen X son independientes, entonces las probabilidades condicionales son de la forma:

$$\mathsf{P}(Y = y\,|X = x) = \Gamma(\varphi(X, \eta) = y). \tag{4.6}$$

En particular, si el ruido es Gaussiano, las probabilidades condicionales y las probabilidades *a posteriori* se pueden escribir en forma de densidades de Gibbss [41], [147].

En esta memoria en general vamos a tratar con bloques de imagen. Las imágenes totales se procesan de la siguiente manera: Para cada pixel en la imagen extraemos una ventana a su alrededor y la clasificamos de acuerdo a un libro de códigos dado. El resultado puede ser tanto el valor central del vector código o la clase asociada al vector vencedor. Llamamos esta aproximación *VQ Bayesian Filter* (VQ-BF) o Filtrado Bayesiano basado en Cuantización Vectorial.

4.2.El modelo para los bloques de imagen

Consideramos que tenemos un conjunto de representantes de los bloques de imagen, i.e. un libro de códigos (*codebook*) $\Omega^* = \{\omega_i^*; i = 1, .., c\}$ donde cada representante es una imagen de tamaño $d \times d$, $\omega_i = \left(x_s^P\right)_{s \in S^{d\times d}}$ donde la malla $S^{d\times d}$ se define como un vecindario

$$S^{N(d\times d)} = \{s : -d/2 \leq |s| \leq d/2\}. \tag{4.7}$$

Dada una muestra de bloques de imagen

$$\mathbf{y}^b = \left\{y_i^b; i = 1, .., n;\right\} \tag{4.8}$$

con

$$y_i^b = \left(y_s^P\right)_{s \in S^{d\times d}}, \tag{4.9}$$

el libro de códigos es el resultado de la ejecución de un algoritmo de diseño de la cuantización vectorial que trata de minimizar alguna función objetivo determinada. Consideramos que el libro de códigos se diseña para minimizar el error cuadrático medio sobre la muestra:

$$\Omega^* = \min_{\Omega} \left\{ \sum_{i=1}^{n} \left\| y_i^b - \omega_{j(i)} \right\|^2 \right\} \tag{4.10}$$

donde

$$j\left(i\right) = y\left(y_i^b\right) = \arg\min \left\{ \left\| y_i^b - \omega_j \right\|^2 ; j = 1, .., c \right\}. \tag{4.11}$$

Esta minimización corresponde a estimación de máxima verosimilitud de los parámetros de una mezcla de Gausianas de matrices de covarianza identidad, que es el modelo asumido para los bloques de imágenes:

$$\mathsf{P}\left(Y^b = y^b\right) = \frac{1}{c} \sum_{j=1}^{c} \frac{1}{\left(2\pi\right)^{d/2}} e^{-\frac{1}{2}\left\| y^b - \omega_j \right\|^2} = \sum_{j=1}^{c} \Pi\left(x^b = \omega_j\right) \mathsf{P}\left(Y^b = y^b \left| x^b = \omega_j \right.\right). \tag{4.12}$$

Por tanto, la búsqueda del vector código más cercano

$$j\left(y^b\right) = \arg\min \left\{ \left\| y^b - \omega_j \right\|^2 ; j = 1, .., c \right\}, \tag{4.13}$$

corresponde al clasificador MAP del bloque de imagen. Asumimos que las probabilidades *a priori* de las clases de los bloques de imagen son las mismas

$$\Pi\left(\left(x^b = \omega_j^*\right)\right) = \frac{1}{c}, j = 1, .., c. \tag{4.14}$$

La decisión de cuantización vectorial dada por

$$\widehat{x} = \omega^*_{j\left(y^b\right)} \tag{4.15}$$

corresponde a la decisión MAP asumiendo que las probabilidades *a posteriori* de los bloques de imagen x^b son de la forma

$$P\left(x^b = \omega_i^* \left| y^b \right.\right) = \frac{\exp\left(-\frac{1}{2}\left\| y^b - \omega_i^* \right\|^2\right)}{\sum_{j=1}^{c} \exp\left(-\frac{1}{2}\left\| y^b - \omega_j^* \right\|^2\right)}, i = 1, .., c. \tag{4.16}$$

Tanto las probabilidades condicionales como las probabilidades *a posteriori* de los bloques imagen se pueden poner en forma de distribuciones de Gibbs con funciones de energía:

$$H\left(Y^b = y^b \left| x^b = \omega_j^* \right.\right) = H\left(x^b = \omega_i^* \left| Y^b = y^b \right.\right) = \frac{1}{2}\left\| y^b - \omega_i^* \right\|^2. \tag{4.17}$$

La consideración de funciones objetivo para el proceso de diseño del cuantificador vectorial distintas de la distorsión darían lugar a otros modelos probabilísticos.

4.3.El modelo para el VQ-BF

En el filtrado bayesiano basado en la cuantización vectorial (VQ-BF) la imagen no se descompone en bloques. Para cada pixel $s \in S^P$ seleccionamos una ventana alrededor suyo $S_s^{N(d\times d)} \subset S^P$. Cada ventana en la imagen $y_s^b = \left(y_{s'}^P\right)_{s' \in S_s^{N(d\times d)}}$ se procesa independientemente. Llamamos bayesiano a este algoritmo porque los vecindarios juegan el papel de condiciones de contorno para el proceso del pixel. El emparejamiento de patrones involucrado es una forma de dependencia no lineal que recuerda las aproximaciones Markovianas del análisis bayesiano de imágenes clásico [41], [147]. La imagen se procesa de dos formas posibles, en modo filtro o en modo clasificación:

1. En modo filtro el pixel restaurado se estima como el pixel central del vector código asociado a su vecindario:

$$\widehat{x}_s^P = \left(\omega^*_{j\left(y_s^b\right)}\right)_{(0,0)}, s \in S^P \tag{4.18}$$

2. En modo clasificación la clase asociada con el pixel es la del vector código asociado al vecindario del pixel

$$\widehat{x}_s^L = j\left(y_s^b\right), s \in S^P \tag{4.19}$$

Vamos a considerar el modo filtro. Si asumimos que los vecindarios de los píxeles son independientes, las probabilidades posteriores son de la forma:

$$\mathsf{P}\left(x \left| y \right.\right) = \prod_{i=1}^{M}\prod_{j=1}^{N} P\left(x_{(i,j)}^b = \omega^*_{j\left(y_{(i,j)}^b\right)} \left| y_{(i,j)}^b \right.\right) \tag{4.20}$$

y la energía *a posteriori* tiene la forma :

$$H\left(x\,|y\right) = C + \frac{1}{2}\sum_{s}\left\|y_s^b - \omega^*_{j\left(y_s^b\right)}\right\|^2. \tag{4.21}$$

La distribución *a priori* corresponde con la probabilidad conjunta de las ventanas de los píxeles.

$$\Pi\left(X = x\right) = P\left(x_s^b = \omega_s^*; s \in S^P\right) \tag{4.22}$$

y no puede ser puesta en forma de producto. Una aproximación para obtener una distribución con forma de Gibbs sería la asunción de que los píxeles vecinos tienen la misma clase o el mismo valor filtrado si las variaciones de las ventanas que los rodean son pequeñas. Esta es una restricción de suavidad *a priori* que depende del libro de códigos. Expresiones apropiadas para la función de energía *a priori* que incorporan esta restricción de suavidad son

$$H\left(x\right) = \sum_{s,t}\left|s-t\right|\left(x_s - x_t\right)^2\left\|x_s^b - x_t^b\right\|^2 \tag{4.23}$$

o

$$H\left(x\right) = \sum_{s}\sum_{t\in N(s)}\left(x_s - x_t\right)^2\left\|x_s^b - x_t^b\right\|^2 \tag{4.24}$$

Tomando en cuenta la expresión de la función de energía *a posteriori* en la ecuacion 4.5 obtenemos como logaritmo de la probabilidad condicional [147]:

$$\ln \mathsf{P}\left(y\,|x\right) = C - H\left(\widehat{x}\,|\widehat{y}\right) - H\left(x\right) \tag{4.25}$$

Asumiendo la función de energía dada por la ecuación 4.24 como la función de energía *a priori*, llegamos a la siguiente expresión para la log-verosimilitud

$$\ln \mathsf{P}\left(y\,|x\right) \approx -\sum_{s}\left\|y_s^b - \omega^*_{j\left(y_s^b\right)}\right\|^2 - \sum_{s}\sum_{t\in N(s)}\left(x_s - x_t\right)^2\left\|x_s^b - x_t^b\right\|^2 \tag{4.26}$$

Si consideramos que el ruido aditivo es Gausiano, entonces las expresiones de las posibles deformaciones que se deducen de la expresión de la verosimilitud son :

$$Y = \varphi\left(X,\eta\right) = \widehat{\eta} - \left(\sum_{s}\left\|y_s^b - \omega^*_{j\left(y_s^b\right)}\right\|^2 - \widehat{\eta}\right) - \sum_{s}\sum_{t\in N(s)}\left(x_s - x_t\right)^2\left\|x_s^b - x_t^b\right\|^2 \tag{4.27}$$

Esta expresión se puede interpretar como la definición de la habilidad del VQ-BF para corregir deformaciones suaves que involucran a los vecinos del pixel.

4.4.Filtrado con VQ-BF y preservación de bordes

Las distribución de probabilidad *a posteriori* del VQ-BF implica que los bordes en la imagen serán preservados siempre que los vecindarios de los píxeles a los dos lados del borde muestren una variación significativa. Esta condición es muy natural y es la razón de las buenas propiedades de preservación de bordes del VQ-BF. Formalmente,

$$\widehat{x}_s \neq \widehat{x}_t \text{ si } j\left(y_s^b\right) \neq j\left(y_t^b\right), \tag{4.28}$$

(asumiendo que vectores código diferentes tendrán píxeles centrales diferentes). Esto corresponde a la probabilidad asignada por la densidad de probabilidad condicional del vector código $\omega^*_{j\left(y_s^b\right)}$ al espacio externo a su región de decisión $R_{j\left(y_s^b\right)}$

$$\begin{aligned} P\left(\widehat{x}_s \neq \widehat{x}_t \left| y \right.\right) &= 1 - P\left(j\left(y_s^b\right) = j\left(y_t^b\right) \left| y \right.\right) \\ &= 1 - \int_{R_{j\left(y_s^b\right)}} \mathsf{P}\left(y_s^b \left| x^b \right.\right) dx^b.. \end{aligned} \tag{4.29}$$

Debe notarse que dos píxeles con idénticos valres pueden recuperarse como diferentes si sus vecindarios cambian abruptamente. Sin embargo, todas estas probabilidades dependen del entrenamiento del libro de código y de las estadísticas de la imagen.

4.5.Cálculo del codebook usando SOM

Un paso crítico en la aplicación de nuestra aproximación es el cálculo del libro de códigos, puesto que debe ser representativo de las auténticas estadísticas de la imagen. Como se ha discutido previamente, estas estadísticas son las piezas de construcción de los modelos *a priori* que subyacen al VQ-BF. El proceso de estimación es un proceso de cuantización vectorial que se puede realizar con las técnicas discutidas en el capítulo 2. Además, es necesario que este proceso se realize en tiempo real. El significado del tiempo real varía de una aplicación a otra. En el caso del cálculo del flujo óptico la respuesta se exigirá en tiempos del orden de fracciones de segundo. En el caso de la segmentación de la MRI la respuesta puede darse en tiempos del orden de fracciones de minuto. En este contexto, la necesidad de esquemas de entrenamiento que se completen en un único paso sobre la muestra es imperiosa. De ahí la discusión y los trabajos

realizados en el capítulo 3 para estudiar la convergencia en un sólo paso sobre la muestra. En los siguientes capítulos aplicamos el SOM en un único paso sobre la muestra extraida de la imagen asumiendolo como un método robusto [51] para la minimización de la distorsión mediante el descenso de gradiente estocástico. A continuación recordamos los fundamentos del SOM tal como se va a aplicar en los siguientes capítulos.

Sea $\mathbf{X} = \{\mathbf{x}_i; i = 1, .., n;\}$ una muestra de vectores dados por bloques de píxeles extraidos de la imagen bajo estudio, y sea $\mathbf{Y} = \{\mathbf{y}_i; i = 1, .., c\}$ el libro de códigos que estamos calculando. La regla adaptativa a la que se ajusta el SOM tras la presentación de un nuevo input $\mathbf{x}_t$ extraido de la muestra en el instante t es la siguiente:

$$\triangle \mathbf{y}_k(t) = \alpha_{k,t} \cdot V(\mathbf{Y}, k, \mathbf{x}_t, t) \cdot (\mathbf{x}_t - \mathbf{y}_k(t)) \quad k = 1, .., c \tag{4.30}$$

donde $V(\mathbf{Y}, k, \mathbf{x}_t, t)$ es la función de vecindad que depende de la topología definida sobre los índices de los vectores código, $\alpha_{k,t}$ es la velocidad de aprendizaje que decrece a cero en la forma habitual para el algoritmo de gradiente estocástico. Para obtener el aprendizaje en un paso sobre la muestra esta velocidad de aprendizaje sigue la ecuación

$$\alpha_{k,t} = \alpha_0 (1 - \tau_k / n) \tag{4.31}$$

donde n es el tamaño de la muestra. Limitamos el número de iteraciones al tamaño de la muestra. La función de vecindad es de la forma

$$V(\mathbf{Y}, k, \mathbf{x}_t, t) = \begin{cases} 1 & |i - j(\mathbf{x}_t)| \leq v_t \\ 0 & \text{otherwise} \end{cases}, \tag{4.32}$$

el radio v_t del vecindario definido por la función de vecindad sigue la expresión

$$v_t = \left\lceil (v_0 + 1)^{\left(1 - \frac{r}{n} t\right)} \right\rceil - 1 \tag{4.33}$$

para $t < \frac{n}{r}$. Para $t > \frac{n}{r}$ asumimos que el SOM se comporta como el algoritmo competitivo simple y la función de vecindad deviene el criterio duro de pertenencia a los agrupamientos. El factor r regula la velocidad con que se realiza la convergencia funcional del SOM al algoritmo competitivo simple. En otras palabras, especifica el porcentaje de la muestra en el que el vecindario del SOM será no nulo.

4.6. Conclusiones

En este capítulo se presenta el algoritmo de filtrado que va a ser utilizado en capítulos posteriores. Partimos de la formulación bayesiana del proceso de la imagen. En nuestro caso, la estimación de máxima probabilidad *a posteriori* (MAP) corresponde a la decisión de asociar un bloque representante al pixel o al bloque que está siendo utilizado. Si asumimos que la distribución de probabilidad *a posteriori* es de Gibbs, su logaritmo corresponde a la función de energía que se descompone en las funciones de energía *a priori* y condicional. A partir de la formulación bayesiana asumiendo un modelo de función de energía *a priori* que incorpora las habituales restricciones de suavidad podemos deducir una cierta forma de la función de energía correspondiente a la d.d.p. condicional. El papel de la d.d.p. condicional habitualmente es el de modelar el ruido aditivo y las deformaciones de la imagen asumidas. Siguiendo este razonamiento, la deducción de la función de energía condicional nos indica que deformaciones de la imagen se corrigen o recuperan mediante la estimación MAP. En el caso del VQ-BF obtenemos la restauración de deformaciones de los vecindarios de los píxeles.

5. Cálculo del flujo óptico

El cálculo del flujo óptico es un problema central en muchas de las tareas involucradas en la aplicación de la visión artificial a la robótica [63], y también en algoritmos de compresión de video. En el cálculo del flujo óptico se puede distinguir entre dos corrientes principales, una basada en el cálculo de las derivadas espacio-temporales y la otra basada en el cálculo de la correlación entre las tramas. Las aproximaciones basadas en las derivadas espacio-temporales [58], [63], [68] son más sensibles al ruido y los efectos de iluminación, y tienden a producir estimaciónes dispersas del flujo óptico. Las aproximaciones basadas en la correlación y el emparejamiento de bloques producen estimaciónes densas del flujo óptico y son mucho más robustas ante el ruido aditivo y los efectos de iluminación. La aproximación basada en la correlación es la preferida en esta memoria.

Proponemos el filtrado de la secuencia de imágenes a través de la aplicación de la cuantización adaptativa del color. La cuantización del color involucra la consideración de vecindarios espaciales para regularizar y suavizar el campo del flujo óptico. Este proceso lo realizamos aplicando el filtrado bayesiano basado en la cuantización vectorial (VQ-BF) presentado en el capítulo 4 en el que cada pixel se codifica de acuerdo con la cuantización vectorial de su vecindario. El libro de códigos se calcula aplicando el SOM [80]. En los experimentos que presentamos más adelante se realiza la cuantización del color de forma adaptativa para cada imagen en la secuencia a partir del libro de código que se había obtenido en la imagen anterior.

Como ya se ha comentado en el capítulo 4 el VQ-BF produce la suavización de las imágenes preservando las fronteras entre objetos. Este tipo de suavización es de gran interés para el cálculo del flujo óptico, reduce el flujo espúreo debido a la iluminación y el ruido aditivo, mientras que la preservación de las fronteras produce estimaciónes mejoradas en los lugares críticos. La calidad de la suavización obtenida con el VQ-BF nos permite basar el cálculo del flujo óptico en la correlación a nivel de pixel entre tramas sucesivas. Los resultados experimentales muestran buenas respuestas con un número de clases de bloques pequeño en el

libro de códigos.

El potencial para la implementación en tiempo real viene de la posible implementación inmediatamente paralela del VQ-BF y de los buenos resultados basados en la correlación a nivel de pixel, que puede ser calculada en tiempo real con hardware convencional. La adaptación del libro de código aplicando el SOM en un solo paso sobre la muestra de la imagen siguiente en la secuencia también podría ser realizada en tiempo real dado el número pequeño de vectores código.

El capítulo está estructurado como sigue: la sección 5.1 se recuerdan algunas aplicaciones del flujo óptico que justifican su interés, en la sección 5.2 se ofrece una revisión de antecedentes en el cálculo del flujo óptico, en la sección 5.3 se presenta el algoritmo de cálculo del flujo óptico basado en el VQ-BF, en la seccion 5.4 se presentan algunos resultados experimentales.

5.1.Aplicaciones del flujo óptico.

Muchas referencias al flujo óptico se refieren al problema clásico de la extracción de información 3D a partir del movimiento (*shape from motion*) [77], [97], [131], en diferentes aplicaciones y circunstancias computacionales. Aquí revisaremos algunas para dar una impresión de las aplicaciones y variedad de los métodos.

La localización de partes en un contenedor se presenta en [142] basado en el flujo óptico calculado como el resultado de emparejar bordes extraidos con un operador Laplaciano de la Gausiana. El seguimiento visual se presenta en [110] usando un método de correlación para el cálculo del flujo óptico. La cámara está montada en un brazo robotizado, la profundidad de los objetos en el sistema de referencia es conocido y la relimentación visual se usa para estimar la distancia entre el efector y el objeto. El conocimiento de los parámetros de movimiento se usa para mejorar el rendimiento de la estimación del flujo óptico. Un trabajo similar [57] informa el uso de la relimentación visual para solventar la imposibilidad de realizar la calibración de la cámara en tareas de alcanzar objetos en el espacio.

El uso del flujo óptico para la navegación de vehículos en exteriores (outdoor) en carreteras se describe en [44] haciendo un énfasis especial en las dificultades computacionales y el uso de información adicional. Emplearon un algoritmo basado en la correlación. En el caso de la navegación de interiores, se presenta un sistema en [102] en el que la combinación de imágenes polares y el flujo óptico se usa para la detección de obstáculos y la navegación. El flujo óptico se usa para estimar el tiempo hasta el impacto y se calcula de forma dispersa sobre lineas que consitutuyen características significativas de la imagen.

Otras aplicaciones del flujo óptico se refieren al reconocimiento de expresiones faciales [30], el reconocimiento de gestos humanos [9], [152]. En [152] se propone un método multiescala que permite el reconocimiento de movimientos a muy diversas velocidades y escalas de tiempo. Para terminar este breve paseo en el campo de las aplicaciones del flujo óptico, referimos la aplicación al análisis del movimiento del corazón en imágenes de resonancia magnética (*MRI*) [106].

5.2.Revisión de antecedentes

El flujo óptico se define [63] como "el movimiento aparente de los patrones de intensidad" en las secuencias de imágenes. Este movimiento aparente se toma usualmente como una estimación del flujo en la verdadera imagen correspondiente a la proyección en el plano imagen del movimiento de los objetos en la escena. El cálculo del flujo óptico es el problema de estimar un campo de desplazamiento que transforma los patrones de intensidad de una imagen a la siguiente. Hay tres aproximaciones básicas [132] a este problema:

1. Los métodos variacionales basados en la formulación diferencial de la restricción de intensidad.

2. Los métodos que calculan la corrlación entre bloques de píxeles de imágenes consecutivas.

3. Los métodos basados en el formalismo Bayesiano y los algoritmos estocásticos de minimización de la energía.

Aunque solo ha sido explicitado para la primera clase de métodos, se puede afirmar que todos los métodos implementan las mismas restricciones fundamentales (brillo, suavidad del campo, etc.) desde diversos puntos de vista computacionales. En términos generales, el cálculo del flujo óptico y el intento de inferir movimientos de los objetos en la escena es un problema mal planteado. El denominado "problema de la apertura" se manifiesta en distintas formas dependiendo de los métodos computacionales escogidos. Es la expresión de la indeterminación inherente al problema.

5.2.1.Restricción de brillo

Sea $E(x, y, t)$ la intensidad de un punto en una imagen t dentro de una secuencia de imágenes. La restricción de brillo [63] se puede formular como

$$\frac{dE}{dt} = 0. \tag{5.1}$$

Esta ecuación significa que la intensidad de una imagen a lo largo del movimiento en la escena debe permanecer constante cuando la iluminación y la estructura de la escena son constantes. Esto nos lleva a la expresión

$$\frac{\partial E}{\partial x}\frac{dx}{dt} + \frac{\partial E}{\partial y}\frac{dy}{dt} + \frac{\partial E}{\partial t} = 0. \tag{5.2}$$

Esta expresión se puede abreviar todavía más

$$E_x u + E_y v + E_t = 0. \tag{5.3}$$

El flujo óptico viene dado por el par (u, v). A partir de las restricciones de intensidad, el cálculo de (u, v) está infra-determinado y la convención es la selección de la solución a lo largo de la normal al gradiente de la imagen dado por (E_x, E_y). La expresión particular del problema de la apertura en este contexto es que el flujo óptico sólo se puede observar en las direcciones normales a los bordes de la imagen. El efecto es que el flujo óptico calculado es denso pero calculado con más precisión en los bordes de la imagen. Para atacar esta indeterminación, se impone una restricción de suavidad, de forma que el problema deviene la minimización de un funcional de la forma:

$$\int\int \left[\left(\left(u_x^2 + u_y^2\right) + \left(v_x^2 + v_y^2\right)\right) + \lambda\left(E_x u + E_y v + E_t\right)^2\right] dxdy \tag{5.4}$$

Debido a la utilización de los gradientes de la imagen, esta aproximación es muy sensible al ruido. Ha habido un número de variaciones y mejoras a este esquema básico. Algunos autores proponen el uso de múltiples restricciones [134] para sobre-restringir el problema y obtener estimaciónes del flujo más robustas. Las restricciones adicionales pueden provenir de combinaciones lineales de los gradientes de la imagen que expresan restricciones de flujo. Sin embargo, se ha encontrado que la validez de la aproximación está condicionada por la estructura de la función intensidad de la imagen. El conocimiento de los parámetros de movimiento de la cámara se usa a menudo en aplicaciones de robótica [142] para mejorar la estimación del flujo óptico. En [15] una restricción de trayectoria que involucra la minimización sobre un número de tramas se añade a las restricciones de intensidad para mejorar la robustez.

5.2.2.Cálculo del flujo óptico mediante correlación

El cálculo del flujo óptico basado en la correlación de trozos de la imagen entre tramas de la secuencia de imágenes se ha presentado en [14], en [58], [68] se presenta bajo el contexto del emparejamiento de imágenes. Se refieren a él a veces como el método de mínimo error cuadrático. Ha sido aplicado ampliamente en algoritmos de compresión de video para los módulos de estimación y de compensación del movimiento.

El algoritmo básico consiste en el emparejamiento de cada pixel y su vecindario de tamaño $v \times v$ en de la siguiente trama dentro de un radio de máximo movimiento η. Para cada pixel (x, y) y desplazamiento (u, w) dentro del radio de movimiento permisible $|u|, |w| \leq \eta$, el emparejamiento que da la mayor verosimilitud de que el pixel se está moviendo a lo largo del vector de movimiento (u, w) al pasar de la imagen $E_t(x, y)$ a la imagen $E_{t+1}(x, y)$ es de la forma

$$M_{t,t+1}(x, y; u, w) = \sum_{i,j=-v}^{v} \phi\left(E_t(x+i, y+j), E_{t+1}(x+i+u, y+j+w)\right), \quad (5.5)$$

donde

$$\phi(a, b) = |a - b| \quad (5.6)$$

es la función de emparejamiento que hemos utilizado, aunque otras como

$$\phi(a, b) = (a - b)^2 \quad (5.7)$$

$$\phi(a, b) = a \cdot b \quad (5.8)$$

son aceptables como funciones de emparejamiento. El vector de movimiento en el pixel (x, y) : $F_{t,t+1}(x, y)$ se determina como

$$F_{t,t+1}(x, y) = \underset{-\eta \leq u, w \leq \eta}{\arg\min} \left\{M_{t,t+1}(x, y; u, w)\right\}. \quad (5.9)$$

Este algoritmo es menos sensible al ruido, pero su complejidad crece cuadráticamente con el tamaño de los bloques de píxeles y el radio del movimiento máximo permitido. El problema de la apertura se manifiesta en las limitaciones y ambiguedades impuestas al proceso de emparejamiento por la restricción en el tamaño de los bloques de píxeles y el radio de movimiento máximo. La aproximación basada en la correlación es, en esencia, un proceso de emparejamiento de texturas, de forma que los objetos en movimiento de intensidad constante mayores que el bloque de píxeles considerado no se perciben como una entidad, y sólo se percibe

el movimiento en las fronteras. Esto es aceptable para aplicaciones de compresión, pero para aplicaciones de visión por computador implica la imposibilidad de reconocer apropiadamente algunas instancias de movimiento.

El algoritmo de correlación de regiones

Supongamos que las imágenes están descompuestas en componentes conectados que corresponde a regiones de pixels de cada una de las clases ω. Estos componentes conectados se pueden representar como imágenes binarias $E_{t,\omega,k}(x,y)$ enumeradas por k. Ahora, ϕ es una función de emparejamiento que devuelve un valor proporcional al emparejamiento de dos imágenes binarias correspondientes a las regiones de la imagen. El emparejamiento que nos da la verosimilitud de que la región conectada $A(x,y) \equiv E_{t,\omega,k}(x,y)$ para algún (t,ω,k) moviendose según el vector (u,w) al pasar de la imagen $E_t(x,y)$ a la imagen $E_{t+1}(x,y)$ es de la forma

$$M_{t,t+1}(A;u,w) = \sum_k \phi\left(A(x,y), E_{t+1,\omega,k}(x+u, y+w)\right), \tag{5.10}$$

donde $\phi(A,B) = A \cdot B \approx \sum_{x,y} |A(x,y) - B(x,y)|$. El vector de desplazamiento $F_{t,t+1}(x,y)$ se determina como

$$F_{t,t+1}(x,y) = \underset{-\eta \le u,w \le \eta}{\arg\min} \left\{ M_{t,t+1}(x,y;u,w) \right\}. \tag{5.11}$$

Esta forma de atacar el problema soluciona parcialmente el problema de la apertura, puesto que no considera bloques de tamaño fijo, sino regiones de la imagen de tamaño y forma variables identificadas mediante algoritmos de segmentación preferentemente no supervisados, como es el VQ-BF discutido en el capítulo 4.

5.2.3.Métodos Bayesianos

Los métodos bayesianos aplicados al cálculo del flujo óptico se han propuesto en [82]. Se basan en la estimación Maxima A Posteriori (MAP) del campo de desplazamiento entre imágenes consecutivas. Este estimador está dado por la maximización de las probabilidades *a posteriori*:

$$P[D = d \,|E_{t+1} = e_{t+1}; e_t] = \frac{P[E_{t+1} = e_{t+1} \,|D = d; e_t]\, P[D = d; e_t]}{P[E_{t+1} = e_{t+1}; e_t]}, \tag{5.12}$$

donde D es el campo de desplazamiento aleatorio, equivalente al par (u, v) referido previamente. La estimación del flujo óptico deviene la búsqueda del campo de desplazamientos que maximiza estas probabilidades:

$$\widehat{d} = \arg\max_{d} \left\{ P\left[D = d \,|\, E_{t+1} = e_{t+1}; e_t \right] \right\}. \tag{5.13}$$

Modelar el flujo óptico consiste en definir las distribuciones de probabilidad *a priori* y condicionales. Las probabilidades condicionales modelan como se produce la imagen a partir de una previa y de un campo de desplazamiento. Esta distribución implementa la restricción de brillo. Es usualmente un modelo del ruido aditivo y se considera gausiano. La distribución de probabilidad *a priori* modela el campo de desplazamiento e implementa las restricciones adicionales que regularizan el problema. Si podemos expresarla como una distribución de Gibbs

$$P\left[D = d; e_t \right] = \frac{1}{Z} \exp\left(-H\left(d \right) \right), \tag{5.14}$$

con

$$H\left(d \right) = \sum_{c \in C} V_c\left(d_c \right) \tag{5.15}$$

entonces la estimación MAP se convierte en la minimización de la función de energía $H\left(d \right)$ que caracteriza el campo de desplazamiento como un Campo Markoviano Aleatorio (Markov Random Field) definido por los potenciales $V_c\left(d_c \right)$ sobre el conjunto de cliqués C que define la estructura de vecindarios en el campo de desplazamiento. Los potenciales son una descripción local de las propiedades buscadas por la minimización de la función de energía. La restricción de suavidad, por ejemplo, se expresa como:

$$H\left(d \right) = \sum_{(x,y) \in C} \left| d\left(x \right) - d\left(y \right) \right|^2. \tag{5.16}$$

La preservación de bordes en la imagen se puede modelar de la misma manera, así como muchas otras restricciones. Estas expresiones se pueden combinar linealmente para implementar las restricciones más sofisticadas que modelan nuestro conocimiento *a priori* sobre las imágenes. El mayor atractivo de esta forma de trabajo es la facilidad con la que se introducen nuevas restricciones y el modelado local de las restricciones. La principal desventaja radica en las dificultades asociadas con la minimización de la función de energía. Los algoritmos de minimización locales dan resultados inapropiados y los métodos de minimización global necesitan tiempos de cálculo muy altos.

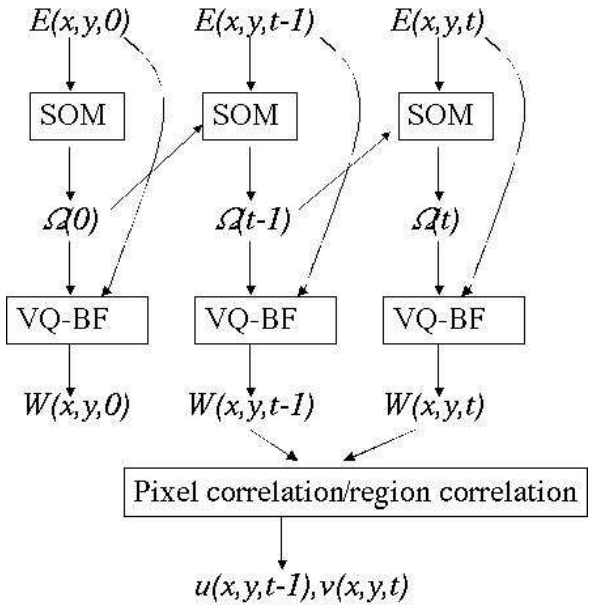


Figura 5.1: Esquema del calculo del flujo óptico basado en el preproceso mediante VQ-BF

5.3.El procedimiento completo

El procedimiento completo para el cálculo del flujo óptico basado en el preprocesado mediante VQ-BF se muestra en la figura 5.1. Dada una secuencia de imágenes $E(x, y, t)$, cálculamos el primer libro de códigos $\Omega(0)$ para VQ-BF de la primera imagen usando el SOM. Este libro de códigos está adaptado a las nuevas imágenes aplicando SOM a la nueva imagen tomando como libro de códigos inicial el correspondiente a la imagen anterior. Por tanto $\Omega(t)$ se calcula a partir de $\Omega(t-1)$ y la imagen $E(x, y, t)$. El algoritmo VQ-BF usando el libro de códigos $\Omega(t)$ se aplica a la imagen $E(x, y, t)$ para obtener $W(x, y, t)$: la clasificación de los píxeles basada en sus vecindarios. Las imágenes resultantes de la clasificación se usan para calcular el flujo óptico (u, v) usando la correlación de bloques o de regiones.

5.3.1.El SOM y el cálculo del flujo óptico

Un paso crítico de nuestra aproximación es el cálculo del libro de códigos, puesto que debe ser representativo de las verdaderas estadísticas de la imagen. También,

el cálculo se debe realizar en forma robusta y rápida. De cada imagen de la secuencia extraemos de forma aleatoria una muestra constituida por un número de bloques de píxeles que pueden solaparse.

Como ya se ha comentado anteriormente, aplicamos un SOM en un único paso sobre la muestra [51] para realizar estos requerimientos. En los experimentos descritos a continuación hemos ajustado la velocidad de aprendizaje como

$$\alpha_{i,t} = (0.001)^{t/t_{\max}} \tag{5.17}$$

con $t_{\max}$ igual al tamaño de la muestra. La función de vecindad es de la forma

$$V_{i,t}(y_t, \Omega_t) = 1/\left(1 + \|y_t - \omega_i\| / v_t\right). \tag{5.18}$$

El radio v_t de la función de vecindad sigue la expresión

$$v_t = (0.001)^{t/t_{\max}}. \tag{5.19}$$

La muestra se presenta sólo una vez para permitir que la ejecución sea lo más cercana posible a las restricciones de tiempo real. Estas restricciones son muy críticas debido a los largos tiempos de convergencia que son característicos de las redes neuronales competitivas como ya se ha comentado en capítulos anteriores.

5.4.Experimentos y resultados

Hemos aplicado nuestra aproximación a algunas secuencias de imágenes con las siguientes precisiones:

- El SOM se aplica a la primera imagen de la secuencia para obtener el libro de códigos inicial.
- El algoritmo competitivo simple (equivalente al SOM con función de vecindad nula) se usa a continuación para realizar pequeñas adaptaciónes del libro de códigos en cada imagen.
- Las imágenes en la secuencia son filtradas con el VQ-BF usando el libro de códigos calculado para ellas de forma específica.
- El flujo óptico se calcula usando la correlación entre bloques de píxeles y de regiones. Esto es, calcula el desplazamiento de cada pixel considerandolo

dentro de un bloque y dentro de una región. En realidad el algoritmo estandard de correlación se calcula con radio de vecindario $v = 0$ y máximo radio de movimiento $\eta = 5$. El algoritmo de correlación de regiones se aplica sobre los componentes conectados identificados por VQ-BF, con el mismo radio de movimiento máximo.

Hemos probado varios tamaños de libros de códigos y dimensiones de los vectores código para el VQ-BF. Las figura en esta sección muestran los resultados para un libro de códigos con 4 vectores y en el que los vectores código son bloques de 5×5 píxeles. Los tiempos de respuesta de un prototipo realizado en IDL© son del orden de segundos. Pensamos que una implementación en C podría llevar estos tiempos al orden de décimas de segundo para cada trama.

En general la suavización producida por el VQ-BF provoca estimaciones nulas del flujo óptico en superficies constantes o casi constantes con alguna microtextura, como las paredes o las puertas, cuando aplicamos la correlación de los píxeles para estimar los vectores de movimiento. Aunque no se realiza ningún proceso de detección de bordes, los bordes significativos son los elementos fundamentales en el flujo resultante. Cuando aplicamos la correlación de regiones, se obtiene una estimación densa y consistente. Mostramos resultados sobre secuencias de panning, zooming y personas actuando.

La figura 5.2 muestra por filas las imágenes originales, los resultados del filtrado VQ-BF, la estimación densa del flujo óptico obtenida mediante el algoritmo de correlación de regiones y las estimaciónes dispersas resultantes de la correlación a nivel de pixel, ambas calculadas sobre las imágenes filtradas con VQ-BF. Se puede apreciar que el SOM encuentra buenas estimaciónes del libro de códigos que conducen a particiones robustas de la imagen en regiones consistentes a lo largo de la secuencia. Una dificultad con nuestra aproximación es la partición de regiones suaves pero grandes debido a la tendencia natural de los algoritmos de agrupamiento a concentrar los representantes de los clusters en regiones del espacio con alta densidad de muestras. Este efecto aparece en la secuencia de panning en la forma en que la pared se parte en dos o más regiones. El algoritmo de correlación de píxeles encuentra bordes espúreos y estimadores de flujo en las fronteras de esas falsas regiones, sin embargo el algoritmo de correlación de regiones encuentra que el movimiento asociado con estas regiones es el mismo y, por tanto, las presenta como una única región en movimiento.

En la figura 5.3 las imágenes mostradas corresponden a una secuencia de zooming, lo que es equivalente a una traslación en el eje Z de la cámara. Esto refleja la misma configuración óptica que un agente móvil moviendose a lo largo de la

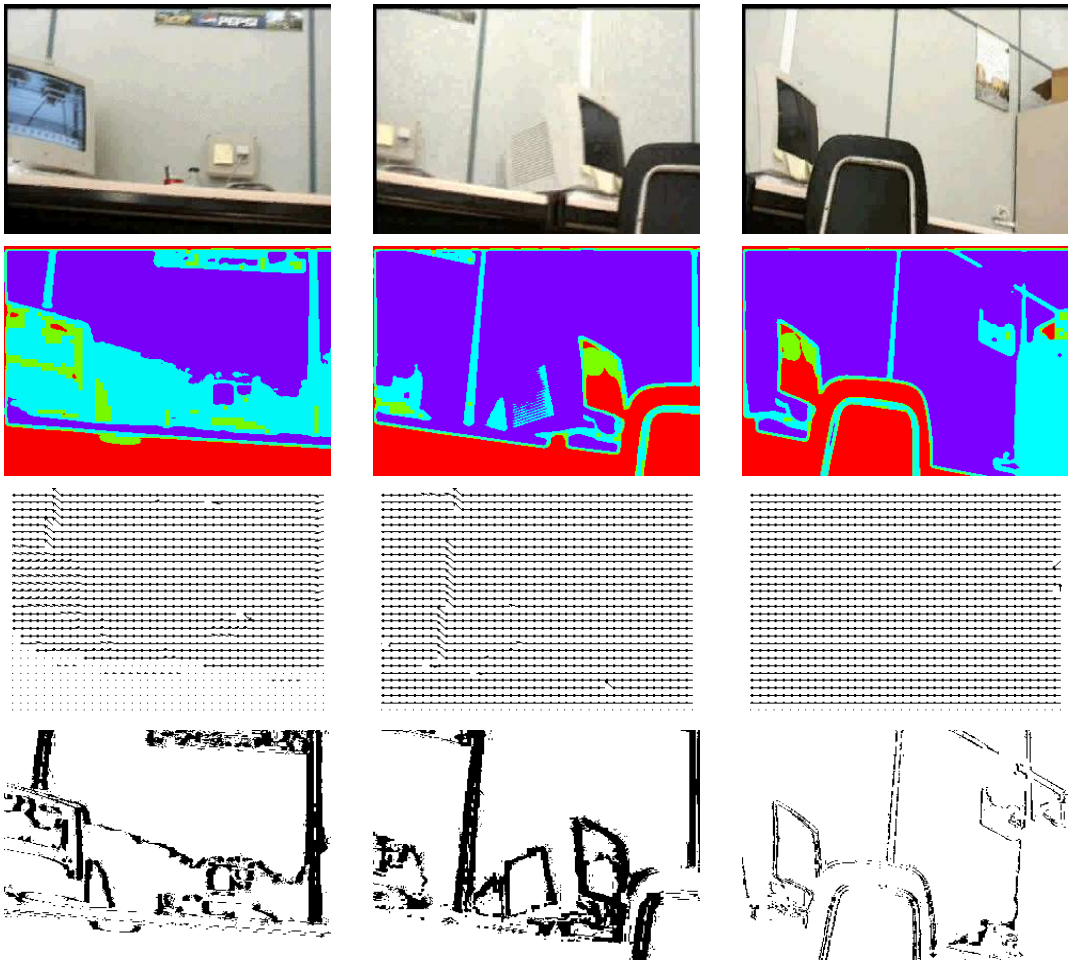


Figura 5.2: Imágenes originales y procesadas en la secuencia de panning (inicial, media, final). Por filas: original, procesada con VQ-BF con 4 clases, estimación densa del flujo óptico basada en la correlación de regiones y la estimación dispersa del flujo basada en la correlación de píxeles

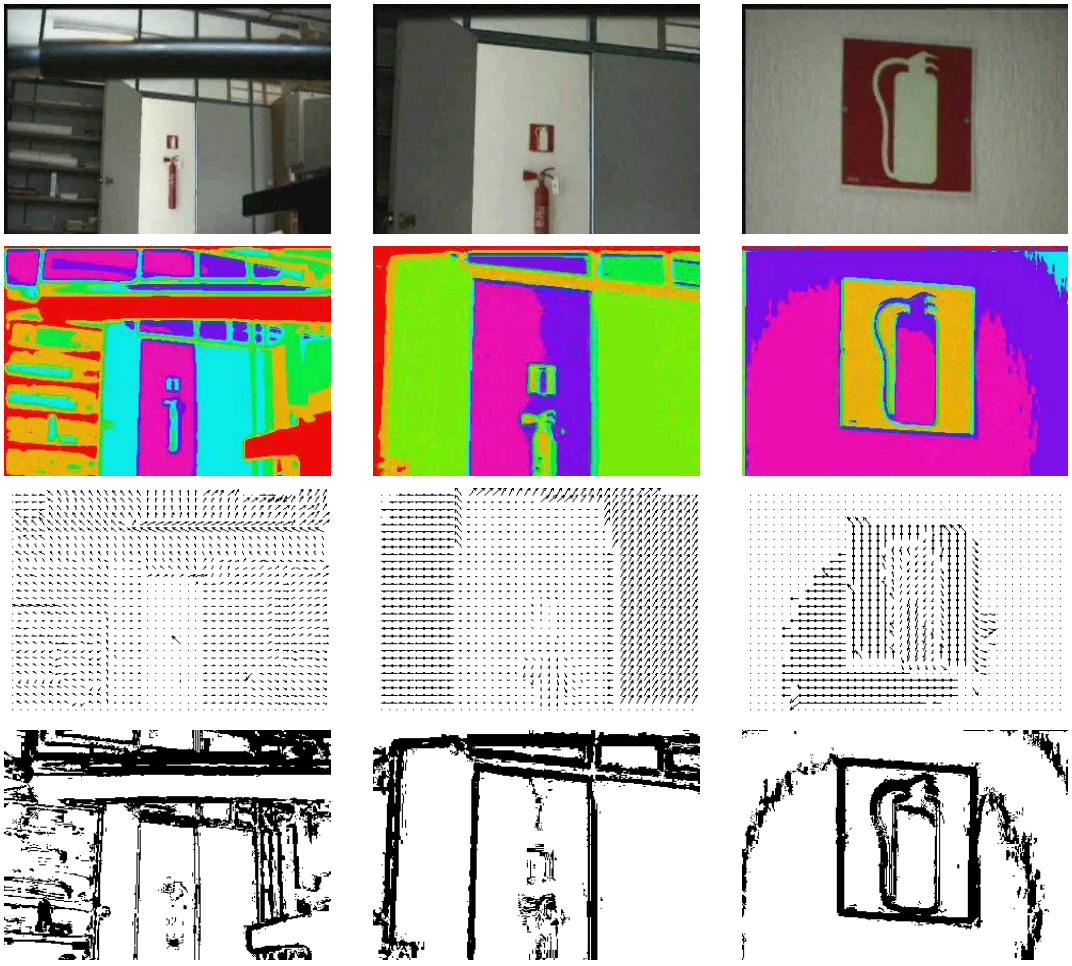


Figura 5.3: Imágenes originales y procesadas en la secuencia de zooming (inicial, media, final). Por filas: original, procesada con VQ-BF con 4 clases, estimación densa del flujo óptico basada en la correlación de regiones y estimación dispersa del flujo basada en la correlación de píxeles

dirección de visualización de la cámara. El flujo óptico disperso resultante del algoritmo de correlación de píxeles muestra los componentes de movimiento fuertes que se detectan en los bordes de los objetos que están en el campo de visión. La estimación densa obtenida por la correlación de regiones muestra también los campos de movimiento consistentes con el movimiento de la cámara. La estimación densa obtenidda es muy suave. Hacia el final de la secuencia hay grandes regiones suaves de la imagen con fuertes gradientes de iluminación. Como en el caso de la secuencia de panning, el algoritmo VQ-BF tiende a partir estas regiones produciendo detecciones espúreas de bordes y algunas estimaciónes de flujo óptico contradictorias.

Finalmente, en la figura 5.4 las imágenes mostradas correspoden a una secuencia en la que dos personas se encuentran y estrechan sus manos. El movimiento de las personas induce cambios en la iluminación de la pared en el fondo, y el algoritmo VQ-BF produce regiones espúreas en la pared. El algoritmo de correlación de píxeles detecta bordes espúreos de flujo en la pared, debido a oscilaciones mínimas de las fronteras de las regiones espúreas. Sin embargo el algoritmo de correlación de regiones no detecta movimiento en la pared del fondo de la escena en la mayor parte de las imágenes. El movimiento de los individuos se detecta bien en la mayor parte de las imágenes.

5.5.Conclusiones y vias de trabajo futuro

En este capítulo hemos descrito la aplicación del algoritmo VQ-BF descrito y analizado en el capítulo 4 al cálculo del flujo óptico. El esquema de cálculo del flujo óptico está basado en la correlación entre las regiones detectadas por el algoritmo VQ-BF. La estimación del libro de códigos se ha realizado con el SOM en un sólo paso sobre la muestra de cada imagen, obteniendo tiempos de respuesta que se aproximan a las necesidades de tiempo real. Las imágenes procesadas con VQ-BF son muy suaves y muestran buenas características de preservación de los bordes en la imagen. La correlación de regiones realizada sobre estas imágenes proporciona una estimación densa del flujo óptico. La correlación a nivel de pixel proporciona estimaciónes dispersas del flujo sobre los contornos de las regiones detectadas por el VQ-BF que pueden corresponder a los objetos en la escena, aunque en algunos casos VQ-BF produce regiones espúreas que no se corresponden con elementos de la escena real. Este es el caso con grandes superficies suaves, como paredes de una habitación, donde el gradiente suave de iluminación se convierte en varias regiones. La correlación a nivel de pixel da detecciones espúreas a los largo de

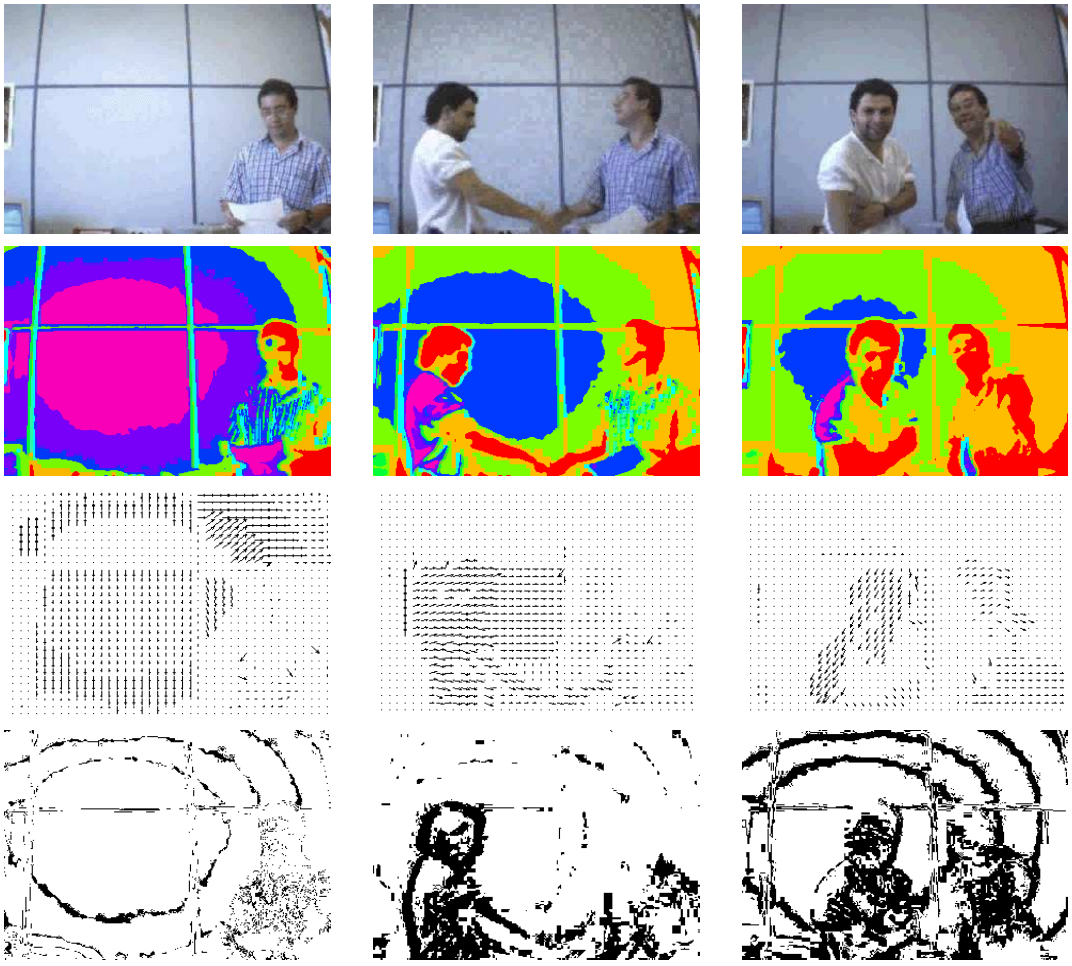


Figura 5.4: Imágenes originales y procesadas en la secuencia de "gente" (inicial, media, final). Por filas: original, procesada con VQ-BF con 4 clases, estimación densa del flujo óptico basada en la correlación de regiones y estimación dispersa del flujo basada en la correlación de píxeles

los falsos bordes de estas regiones. Sin embargo, la correlación a nivel de región agrupa estas particiones fantasma. La combinación de ambas estimaciónes da lugar a estimaciónes robustas y precisas del flujo óptico.. La acumulación del flujo [146] puede ser de ayuda para detectar inconsistencias en los fujos atribuidos a regiones espúreas producidas por los efectos de la iluminación.

Como líneas de trabajo futuro, es necesario dedicar esfuerzos a la determinación del número apropiado de vectores código. Tembién es importante la detección de cambios bruscos globales de la escena en base al error de cuantización. Estos cambios globales deberían forzar la reinicialización de todo el proceso de cálculo del libro de códigos.

6. Segmentación de imagen de Resonancia Magnética Nuclear usando VQ-BF

En este capítulo se recogen algunos resultados de la aplicación del VQ-BF en las segmentación de imágenes 3D de resonancia magnética nuclear (MRI). Basicamente estos resultados han sido publicados en [114] y se obtuvieron en colaboración con los miembros del IRM de la UCM, actual Instituto de Estudios Biofuncionales. El papel que juega el VQ-BF es el de un preproceso no supervisado, basado en la propia información de la imagen procesada, que permite mejorar los resultados de segmentanción supervisada obtenidos por un clasificador construido mediante una red neuronal artificial clásica, la red de alimentación hacia delante entrenada con el algoritmo de retropropagación del gradiente, también denominado Perceptrón Multicapa (MLP). El sistema propuesto es, en su esquema general, un sistema híbrido con una capa construida mediante entrenamiento no supervisado y otra construida mediante entrenamiento supervisado. El MLP se entrena sobre un único corte del volumen, que contiene la zona central de la región a extraer, para luego generalizar la clasificación a todo el volumen 3D para extraer el volumen correspondiente al tejido o estructura objetivo.

El método se ha aplicado a imágenes de resonancia magnética nuclear recogidas en un experimento de modelo de un proceso inflamatorio agudo (pesadas en T_2) y de un estudio clínico de la enfermedad humana de Alzheimer (imágenes pesadas en T_1). En el primer caso se obtiene una alta correlación de las medidas volumétricas del tejido infectado y el tejido sano entre las proporcionadas por la segmentación manual, por la proporcionada por el método propuesto, y por los estudios histopatológicos. Los resultados sobre el estudio clínico fueron similares en la medición volumétrica del hipocampo y del *corpus callosum*.

En la sección 6.1 se realiza un breve estudio del estado del arte. En la sección 6.2 se presentan las imágenes que se han utilizado en la validación de nuestra aproximación. En la sección 6.3 se presenta el método semi-automático prop-

uesto. En la sección 6.4 se presentan los índices estadísticos utilizados para validar nuestro algoritmo. En la sección 6.5 se presentan los resultados cuantitativos obtenidos sobre los datos experimentales. Finalmente, la sección 6.6 se presentan las conclusiones sobre la aplicación del método semi-automático de segmentación.

6.1.Revisión del estado del arte

Las técnicas de imagen médica actuales han demostrado una alta fiabilidad para la diagnosis diferencial y para la evaluación de la respuesta a las terapias de algunas patologías. Entre ellas destaca la imagen de resonancia magnética (MRI) por su flexibilidad y capacidad diagnóstica. Su combinación con técnicas de procesado de imagen digital y de reconocimiento automático de patrones aumenta la precisión en la cuantificación del tamaño de las diferentes lesiones y en la extracción de sus características[70], [81]. Esto resulta en una reducción del tiempo de análisis, una reducción del sesgo debido al operador y la identificación consistente de tipos de tejidos en distintas imágenes. Un objetivo a largo plazo es establecer una metodología automatizada y precisa para realizar la segmentación y la medición volumétrica de las regiones en las imágenes. Además, los métodos no subjetivos son especialmente útiles cuando la decisión se debe tomar por consenso entre varios médicos [70]. En este sentido, las redes artificiales neuronales (ANN) [10], [62], [113] y los métodos de reconocimiento estadístico de patrones [28], [7] son herramientas apropiadas para construir sistemas de análisis automatizado de las imágenes médicas. Las ANN han sido reconocidas como herramientas de ayuda a la decisión [70], que permite construir clasificadores basados en características cuantitativas y cualitativas extraidas de las imágenes médicas: i. e., diagnosis en mamografías [151] y segmentación de la estructura del cerebro [93], [46]. La segmentación de la imagen médica se realiza por la clasificación de los píxeles de la imagen, asociando cada pixel a una estructura o tejido concreto. Otras aproximaciones tratan de detectar los contornos de las regiones sin asociar tejidos y píxeles explícitamente.

Un problema intrínseco a todos estos métodos completamente automatizados es su capacidad final para tratar con formas complejas y la variabilidad en los tamaños y formas de los tejidos. Los tejidos vivos se deforman de forma no lineal e impredecible, y su respuesta a los sistemas de visualización que generan las imágenes médicas (Rayos X, Resonancia Magnética Nuclear, etc.) puede variar considerablemente dependiendo de las condiciones patológicas o simplemente variables. Por ello, se incluye generalmente una parte interactiva o supervisada en

el proceso de modelado y clasificación para asegurar un resultado más fiable. Por ejemplo, las denominadas serpientes o modelos de formas deformables o activos [98] necesitan la especificación de un contorno inicial razonablemente cercano a la región de interés. Este contorno inicial proviene de un conocimiento *a priori*. Estos métodos proporcionan en algunos casos resultados excelentes aunque el procedimiento general para la inserción por parte del operador humano de marcas necesarias en la imagen es trabajosa, el proceso computacional de ajuste de las superficies es largo y se sabe que es, precisamente, muy sensible a las condiciones iniciales [129].

Los estudios clínicos terapeúticos (p.ej.: volumenes de tumores cerebrales) demandan este tipos de procedimientos automatizados de visualización y de análisis de las imágenes médicas, que permiten la monitorización no invasiva de los procesos de algunas enfermedades y de los efectos de un tratamiento farmacológico sobre la morfología, fisiología o bioquímica de un tejido [86]. Estos procedimientos pueden ayudar a acelerar la evaluación del mecanismo y los perfiles farmacinéticos, farmacodinámicos y de seguridad de una droga candidata en la investigación preclínica mediante modelos animales [124] . Estos procedimientos de análisi permiten la realización de estudios longitudinales sobre el mismo individuo, lo que reduce el coste económico y mejora la precisión experimental puesto que reduce la necesidad de combinar las observaciones sobre distintos individuos de las distintas fases de un proceso, debido a la necesidad de sacrificar el individuo para realizar las observaciones histopatológicas. La reducción consecuente en el número de sujetos experimentales que se requieren es espectacular.

Algunas aproximaciones diagnósticas a algunas enfermedades neurodegenerativas se basan en las medidas volumétricas de las diferentes estructuras en el cerebro humano, p.ej.: se han observado diferencias significativas en los volúmenes del *gyrus* del hipocampo y para-hipocampo en pacientes con enfermedad de Alzheimer y en personas sanas. Además, esas observaciones han sido correlaciónadas con la medición global de algunas funciones cognitivas [126]. Desde otro punto de vista, la determinación de los volúmenes 3D del hipocampo es un caso de estudio excelente para verificar los límites de cualquier nuevo método automatizado de segmentación de imágenes médicas, debido a su pequeño tamaño y ambiguedad en la definición de la estructura [93], [11]. El *corpus callosum* se reconoce como un indicador de la conectividad entre los hemisferios cerebrales. La medición de su volumen también es de interés para el estudio de algunas neuropatologías.

Nuestro procedimiento utiliza el mapa autoorganizativo de Kohonen (SOM) [79], [80] y el MLP. El SOM se ha usado para procesar imágenes de resonancia

magnética (MRI) multi-espectrales[1] y funcionales con el objetivo de obtener el agrupamiento de perfiles de voxeles. El SOM y el MLP se han usado en aproximaciones híbridas, que combinan técnicas supervisadas y no supervisadas, para la detección de patologías, como el osteosarcoma en [115]. Aparte de este método, se encuentran referencias en la literatura sobre la aplicación de otras técnicas como el agrupamiento borrosos a la segmentación de las MRI [19], [127], [4]. La mayor parte de los trabajos en la segmentación de MRI se aplican a imágenes multiespectrales [130], [64], sin embargo, el mayor tiempo de adquisición de las imágenes mutiespectrales y la necesidad de un registro fino entre las imágenes obtenidas con cada secuencia de pulsos justifica la investigación en la segmentación de imágenes monoespectrales obtenidas con una única secuencia de pulsos.

6.2.Las imágenes experimentales

En este capítulo hemos escogido tres problemas importantes y difíciles para aplicar la metodología interactiva de segmentación. Hemos realizado la segmentación, evaluado los volúmenes segmentados comparándolos con los obtenidos manualmente e histológicamente (en ocasiones). Nuestro trabajo intenta la detección de regiones de interés 3D que corresponden a estructuras de tejidos con gran robustez frente a traslaciones y deformaciones, tanto entre rodajas en el volumen 3D como entre diferentes volúmenes MRI de la misma estructura. Se espera que estos métodos sean útiles en aplicaciones clínicas, biológicas y farmacéuticas. En esta sección vamos a detallar las características de las imágenes experimentales sobre las que se han aplicado los algoritmos de segmentación. Se trata de un modelo animal de infección y de dos estructuras del cerebro para las que se han utilizado imágenes clínicas.

6.2.1.Datos del modelo experimental animal

Los estudios seriales de los ratónes (en número de 16) inoculados intramuscularmente con *Aspergillus fumigatus* se realizaron con cada animal en diferentes días de infección aguda, desde los días 0 a 14 tras la inoculación. La visualización se

[1]Las Imágenes de MR se generan mediante la excitación mediante secuencias de pulso electromagnético. Las Imágenes generadas por una sola secuencia las denominamo "monespectrales", las generadas por un conjunto de pulsos diseñados con algún proposito especifico las denominamos "multiespectrales" siguiendo la nomenclatura de las Imágenes de reconocimiento remoto. Cada pixel tiene un vector de valres asociado según esta interpretación.

realizó utilizando un espectrómetro Bruker Biospec 47/40 (Ettlingen, Germany) con un resonador de tipo jaula de pájaros hecha a medida. Los animales se colocaban en posición prona en posiciones similares en cada experimento. Los animales se colocaban de forma que las dos patas se insertan lado a lado en la bobina. Después de una secuencia de exploración, se adquirieron conjuntos de datos 3D pesados en T_2 rápidos ($256 \times 256 \times 32$) de imágenes axiales con valores de TR/TE de 2000/67.5 ms y campo de visión de $40 \times 40 \times 22$ mm. El método propuesto se aplica a la cuantificación del volumen del músculo inflamado y la necrosis en una lesión con abceso con un curso de inflamación agudo y crónico. Sólo las imágenes correspondientes a los días 3, 7, y 14 tras la inoculación se usaron para el validar el proceso semi-automático de proceso y cuatificación de la imagen propuesto, puesto que los correspondientes animales sólo fueron estudiados en esos días. Los detalles del estidio histológico se encuentran en [114]. Se intentó hacer corresponder cada plano de las secciones histológicas con los cortes axiales de MRI.

6.2.2.Datos clínicos

Los detalles del protocolo del estudio clínico se encuentran en [114]. Se segmentó el hipocampo y el *corpus callosum* en cuatro de los sujetos, en la imagen total del cerebro como estructuras candidatas excelentes para detectar diferencias. La MRI se realizó usando un escaner Signa General Electric 1.5T Medical System. Se empleo un resonador estandard General Electric de jaula para la adquisición de las imágenes. Tras una secuencia de exploración, se realizó una adquisición fast-spoiled–gradient-recalled en el estado estacionario se adquirieron conjuntos de datos 3D ($256 \times 192 \times 124$) en el plano axial con valores TR/TE de 14.6/3.1 ms y campo de visión $240 \times 180 \times 160$ mm.

6.3.El procedimiento semi-automático de segmentación

El proceso semi-automatizado aplica dos tipos de redes neuronales (ANN) para aislar y clasificar las estructuras 3D y permitir la medición de su volumen. Los datos originales se preprocesan inicialmente por medio de VQ-BF descrito en el capítulo 4. La subsequente identificación semi-automática de la lesión o la región de interés (ROI) se consigue por la aplicación de un MLP. El resultado de la aplicación de esta segunda ANN es otro volumen con la misma dimensión que el original, donde el valor 1 representa el ROI en el conjunto de datos 3D. En alguna de las aplicaciones, tras seleccionar el ROI, una nueva aplicación no super-

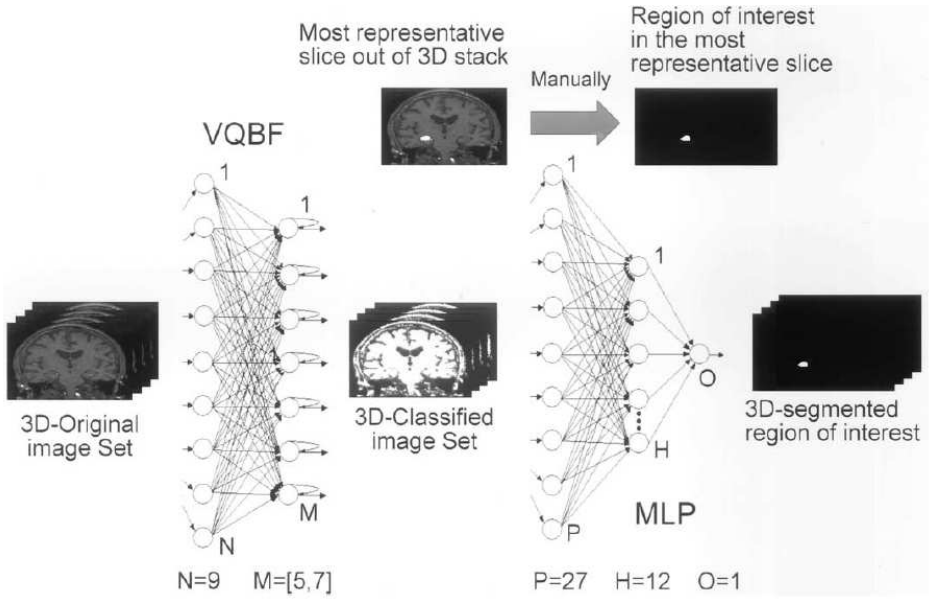


Figura 6.1: Descripcion grafica del procedimiento de segmentacion de imagen MR propuesto

visada de VQ-BF puede ser necesario, p.ej: la distinción de tejidos edematosos y necrotizados en la región inflamada [125]. Sigue una explicación detallada del procedimiento semi-automatico propuesto

6.3.1.Segmentación no supervisada mediante VQ-BF

El algoritmo VQ-BF se puede visualizar como la aplicación de la cuantización vectorial a vecindarios solapados de los voxeles, seleccionados por una ventana deslizante, ver figura 6.1. Los vecindarios de los vóxeles se definen en tres dimensiones puesto que tratamos con datos 3D: la tripleta $(\mathrm{n}X, nY, nZ)$ da el tamaño del vecindario en cada dimensión, donde el eje Z denota el número de corte. Como se describe en el capítulo 4 el VQ-BF se basa en un libro de códigos extraido de la propia imagen, esta estimación se realiza mediante la aplicación del algoritmo SOM descrito en el capítulo 2.El entrenamiento sigue las formulas convencionales del SOM. Hemos aplicado la siguiente función de vecindad específica para cada

unidad

$$v(t) = \frac{\eta(t)(\rho(t)+1)}{d} \tag{6.1}$$

donde $\eta(t)$ es la velocidad de aprendizaje que decrece exponencialmente desde 0.5 a 0.01, $\rho(t)$ es un factor de distancia que también cambia durante el proceso iterativo desde 1 hasta 0.001, finalmente d es la distancia positiva entre la neurona vencedora y la actualizada en el espacio de los índices.

El preproceso consiste en la clasificación de cada voxel de acuerdo a su vecindario, donde el vecindario de un voxel es un cubo centrado en dicho voxel [52], [53], [55]. El resultado es un nuevo volumen 3D con las mismas dimensiones espaciales que el volumen original. El número de niveles de gris dependen del número de clases deseadas para caracterizar el tejido en la imagen MRI.

En el modelo animal experimental, el proceso de entranamiento de VQ-BF se aplicó con los siguientes parámetros: un vecindario de tamaño $3 \times 3 \times 1$ y 3000 iteraciones sobre una muestra de 3000 combinaciones de inputs seleccionadas aleatoriamente. Estos parámetros se seleccionaron para garantizar un tiempo de cálculo razonable. Este tamaño de vecindario preserva las definiciones de las fronteras; y la iteración y el número de combinaciones de input asegura la convergencia de los vectores de pesos a salidas con significado. Los mismos parámetros se usaron para los datos clínicos, excepto que el máximo número de iteraciones permitidas se incrementa a 5000, así como el número de representantes mencionados arriba. Por esta razón en nuestro caso en la capa de entrada $N = 9$ (correspondiendo al tamaño del vecindario) y en la capa de salidad $M = 5$ y $M = 7$ en caso del modelo animal y en el caso clínico, respectivamente, equivalente al número de clases (número de niveles de gris) en el conjunto de imágenes resultantes. Estos números permiten suficiente capacidad para definir múltiples clases de tejidos. La decisión final concerniente al número de clases representativas se tomó bajo la opinión experta del patólogo y el neuroradiologo. En el caso anima, por ejemplo, los representantes se asociaron consistentemente con [125]: fondo, músculo sano, abcesos, músculo inflamado y un grupo de tejidos incluyendo grasa subcutánea o intermuscular o tejidos con una alta señal T2 en la periferia de las lesión. No hemos intentaso la determinación automática del número de clases.

6.3.2.Identificación supervisada del volumen

El MLP es una de las ANN supervisadas más conocidas [62], [113]. Consiste en una red de alimentación hacia delante entrenada con el algoritmo de retropropagación del gradiente. Se aplica a la detección de una ROI en el conjunto

de imágenes que resultan del VQ-BF. En nuestro caso, el MLP consiste de tres capas de unidades computacionales: input, ocultas y output (ver figura **??**). estas capas están completamente conectadas. Las imágenes usadas como input para el MLP son los volúmenes resultantes del procesado con el VQ-BF. La capa de salida se compone de una única neurona binaria. La capa de entrada consiste de P unidades las intensidades de un voxel y sus vecinos, y dos inputs (X_p y Y_p) asociados con la geometría y la posición del voxel.

$$\begin{aligned} X_p &= C_x \left(X_i - X_c\right)^2, \\ Y_p &= C_y \left(Y_i - Y_c\right)^2, \end{aligned} \tag{6.2}$$

donde (X_i, Y_i) son las coordenadas del pixel, y (X_c, Y_c) y (C_x, C_y) son, respectivamente, las coordenadas del centro de masa y las desviaciónes estandard de la región objetivo en la imagen de clasificación deseada obtenida por el etiquetado manual de la imagen. Hemos probado diversas configuraciones de capas ocultasy hemos encontrado que la mejor opción consiste en 12 neuronas por capa. Este valor se determinó empíricamente por prueba y error como el mejor convenio entre el coste computacional, la precisión y habilidad para generalizar el clasificador que resuelve el problema de segmentación. Para el entrenamiento del MLP, el operador humano selecciona el corte más característico de la pila 3D original, y realiza la selección manual de un ROI. La imagen binaria producida por esta segmentación manual se utilizará como la clasificación deseada para el proceso de entrenamiento del MLP. El entrenamiento se realiza en las imágenes correspondientes al corte seleccionado y el MLP entrenado se aplica a los restantes cortes de la pila 3D. El entrenamiento se realiza usando el algoritmo de retro-propagación del gradiente con un factor de momento [62]. Para evitar efectos de saturación no deseados en las funciones de transferencia de las unidades, primeramente ordenamos en orden descendente las etiquetas de los representantes de las clases dados por el algoritmo VQ-BF en la región objetivo del corte seleccionado. La velocidad de aprendizaje es igual a 0.45 y el factor de momento adicional es 0.01. Una función sigmoidea con coeficiente 0.5 se usa como la función de activación en todas las neuronas del MLP, excepto la neurona de salida que es binaria. Además de las coordenadas normalizadas de la ecuacion 6.2, los inputs al MLP consisten en las intensidades de los vóxeles que están en un vecindario de tamaño 5×5 de cada voxel. En cualquier caso, el tamaño del vecindario no es un parámetro crítico puesto que vecindarios de tamaños 3×3 y 7×7 producen resultados similares. El proceso iterativo se detiene cuando el máximo número de iteraciones predefinido se alcanza. Tras el entrenamiento, el MLP se aplica a los cortes restantes de la pila

3D y se asigna un valor de cero o uno a cada voxel. El resultado de la aplicación del MLP en el mismo corte utilizado para el entrenamiento se usó para validar la eficiencia del proceso de entrenamiento.

6.4. Análisis estadístico

Para obetener una referencia objetiva para los resultados de la aplicación del algoritmo semi-automatico, dos investigadores independientes realizaron la segmentación de algunas imágenes del conjunto de datos originales. La separación temporal entre estas segmentaciones nunca es menor de una hora para minimizar el error subjetivo de colocación manual de las fronteras relativo a las referencias anatómicas. Las trazas se dibujan en cada corte en un orden no consecutivo para evitar nuevamente la influencia de la memoria. Segmentaron manualmente por triplicado todas las rodajas de tres de los animales.

Todos los cortes de tres de los animales fueron segmentados manualmente por triplicado. Además, el conjunto completo de cortes originales se segmenta manualmente para todos los animales estudiados con el objetivo de realizar un análisis estadístico detallado de la influencia del corte escogido para alimentar el entrenamiento del MLP en los resultados finales.

Los datos se analizaron por ANOVA o procedimientos de correlación para determinar cualquier diferencia estadísticamente significativa entre las áreas para cada corte de la región de interés (ROI) y los volúmenes para el ROI completo. Para todas las comparaciones, se realizaron tests verificando que tanto las medias (test de Student con $p < 0.01$) como las varianzas (test F de Snedecor con $p < 0.01$) eran estadísticamente iguales.

Sin embargo, la comparación entre resultados manuales y asistidos por computador pueden ser engañosos, puesto que las áreas o volúmenes absolutos calculados por ambos métodos pueden ser similares y no incluir los mismos voxeles y, por tanto, el mismo tejido. Por esta razón, tenemos que realizar un análisis estadístico del método semi-automático propuesto versus la segmentación manual. Este análisis está basado en un análisis simplificado de la curva ROC [99], [156]. Se evalúan los siguientes índices de comportamiento:

1. Areas solapadas o volúmenes (S), también llamada similitud o repetibilidad, se define como sigue [11], [76], [119], [29]:

$$S = \frac{|A \cap B|}{|A \cup B|} \tag{6.3}$$

2. El índice de similaridad de Kappa (K_i) [34–36,16][11], [24], [3], [157] definido como sigue:

$$Ki = \frac{2\,|(A \cap B)|}{|A + B|} \tag{6.4}$$

3. La fracción positiva cierta (TPF), que proporciona una medida de sensibilidad del método correspondiendo a la probabilidad de detección

$$TPF = \frac{|A \cap B|}{|B|} \tag{6.5}$$

4. La fracción de falsos positivos (FPF) que está relacionada con la probabilidad de falsa alarma y da un medida de la especificidad

$$FPF = \frac{|A - B|}{|B^c|} \tag{6.6}$$

6.5.Resultados

En esta sección presentamos primeramente los resultados obtenidos sobre el modelo animal de infección por inoculación de A. Fumigatus. Después recogemos los resultados obtenidos sobre los sujetos humanos.

6.5.1.Datos experimentales sobre animales

La aproximación metodológica seguida en todos los caso se puede apreciar en el ejemplo peresentado en la figura 6.2. La figura 6.2A muestra un corte escogido de la lesión central de la imagen MRI original mientras que la figura 6.2B muestra el resultado de la aplicación de VQ-BF. El área dañada está bien segmentada e incluye algunos tejidos con lesiones. El dibujo manual para clasificar la ROI en áreas afectadas y no afectadas (verdad del terreno) tal como la especifica el especialista para la imagen en la figura 6.2A se muestra en la figura 6.2C. Las imágenes en la figura 6.2B (patrón de entrada) y figura 6.2C (patrón de salida) se usaron para entrenar el MLP que fue aplicado luego a las restantes rodajas del volumen incluida la de entrenamiento. La figura 6.2D muestra la imagen tras la segunda aplicación de VQ-BF sobre las imágenes originales enmascaradas por el resultado de la clasificación con el MLP. El recuento automático de los voxeles de cada clase, por el volumen de cada voxel individual proporciona la medida del volumen de tejido infectado.

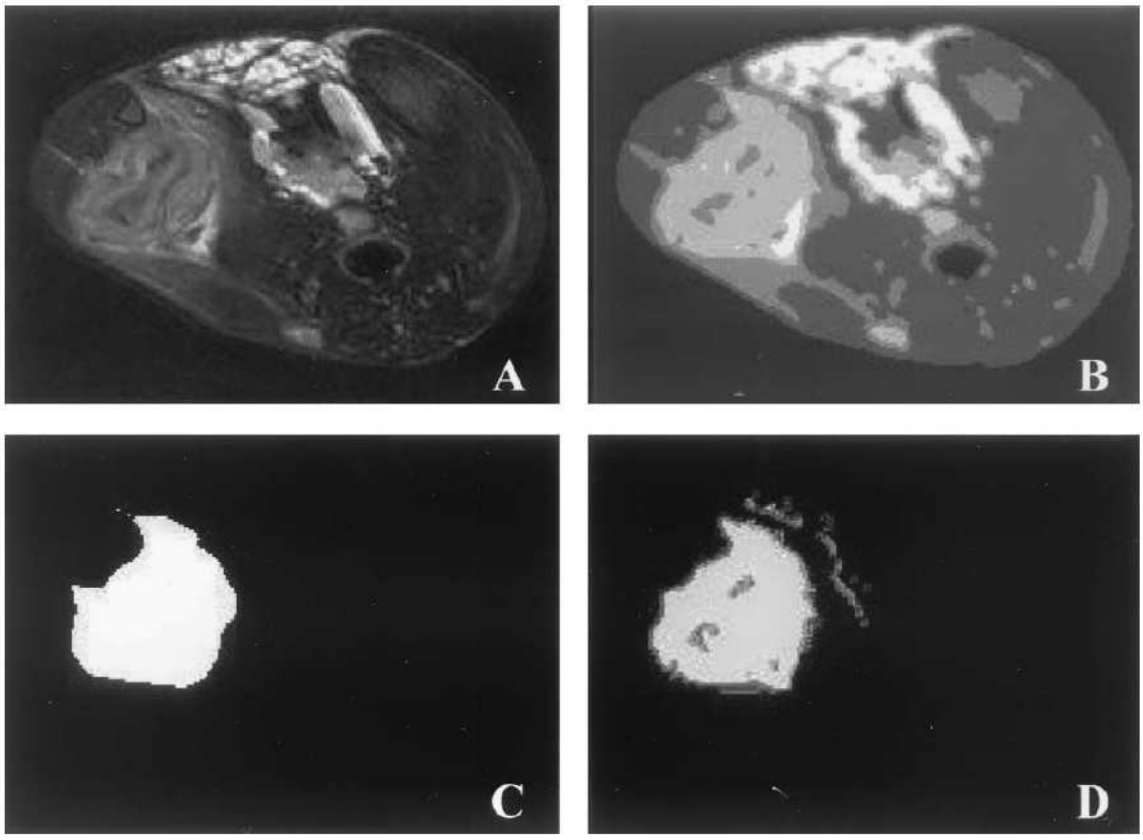


Figura 6.2: Imágenes axiales del ratón tras siete dias de inoculación. A Imagen pesada en T2 original. B Imagen tras la aplicación del VQ-BF. C Imagen objetivo dibujada manualmente indicando la region

Por razones estadísticas, seleccionamos sólo diez rodajas (siempre alrededor de la rodaja central) en tres sujetos. Usamos cada una de ellas alternativamente como la rodaja característica para el entrenamiento del MLP. Se aplica un ANOVA para comparar el método propuesto y el método manual de segmentación en un intento de determinar si las diferencias observadas eran debidas al error aletorio o a diferencias significativas entre los métodos.

No se encontraron diferencias significativas para ninguno delos tres animales seleccionados ($p < 0.05$) tanto entre las áreas medidas usando las dos metodologías como entre los valores obtenidos de los tres intentos manuales. Había, como se podía esperar diferencias estadísticamente significativas entre los valores medidos para los tres ratónes.

El siguiente paso estaba planeado para determinar la sensibilidad del métodos semi-automático propuesto respecto de la rodaja seleccionada para el entrenamiento del MLP. Los resultados de los 10 posibles (uno por cada rodaja) objetivos delineados se dividieron en tres conjuntos correspondientes a: las tres primeras rodajas, las cuatro centrales y las tres finales. La figura 6.3 muestra la comparación del número promedio de píxeles segmentados manualmente para el área inflamada (cuadrádos sólidos) con los determinados por el método automátizado (círculos sólidos). Cada columna de gráficos en la figura corresponde al uso de tres conjuntos (inicial, medio y final) en el proceso de entrenamiento del MLP, y cada fila corresponde a los resultados sobre un ratón. Se usan las tres rodajas centrales para el entrenamiento de los clasificadores en los casos A, D y G, los cuatro centrales en los casos B, E y H o las tres rodajas finales en los casos C, F e I. Los casos A. B y C corresponden a un animal estudiado seis días después de la inoculación con *A Fumigatus*. Los casos D, E y F corresponden a otro animal inoculado siete días antes de tomar la imagen, y los casos G, H e I corresponden a otro animal también estudiado siete días tras la inoculación. Las diferencias entre los valores calculados a partir de la segmentación manual y la calculada por el método semi-automático disminuye cuando el conjunto central de rodajas es el usado para el entrenamiento del MLP, como puede esperarse dado que el error de etiquetado manual en esas rodajas centrales es mucho menor. Una conclusión clara de la figura 6.3 y de los resultados del análisis ANOVA es que el acuerdo entre ambos tipos de medidas se pierde cuando las rodajas que corresponden a los extremos de la lesión se usan para el entrenamiento del MLP. Cuando la mejor rodajas se usa para entrenar el MLP, la mayor discrepancia entre la segmentación manual y la computerizada se observa en las fronteras de la lesión, donde la lesión inflamatoria está menos definida y pobremente delimitada, como se muestra en

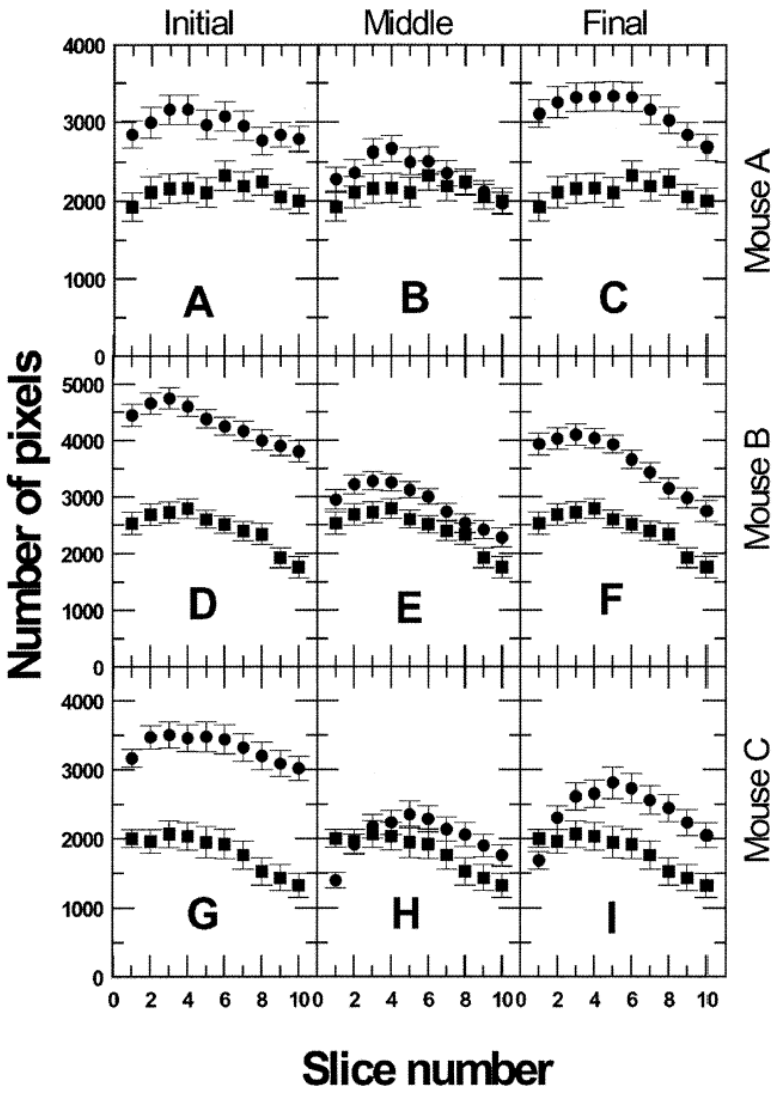


Figura 6.3: Número promedio de píxeles pertenecientes a la zona inflamada para las rodajas indicadas, el cuadrado indica la medida manual y el circulo la medida automática.

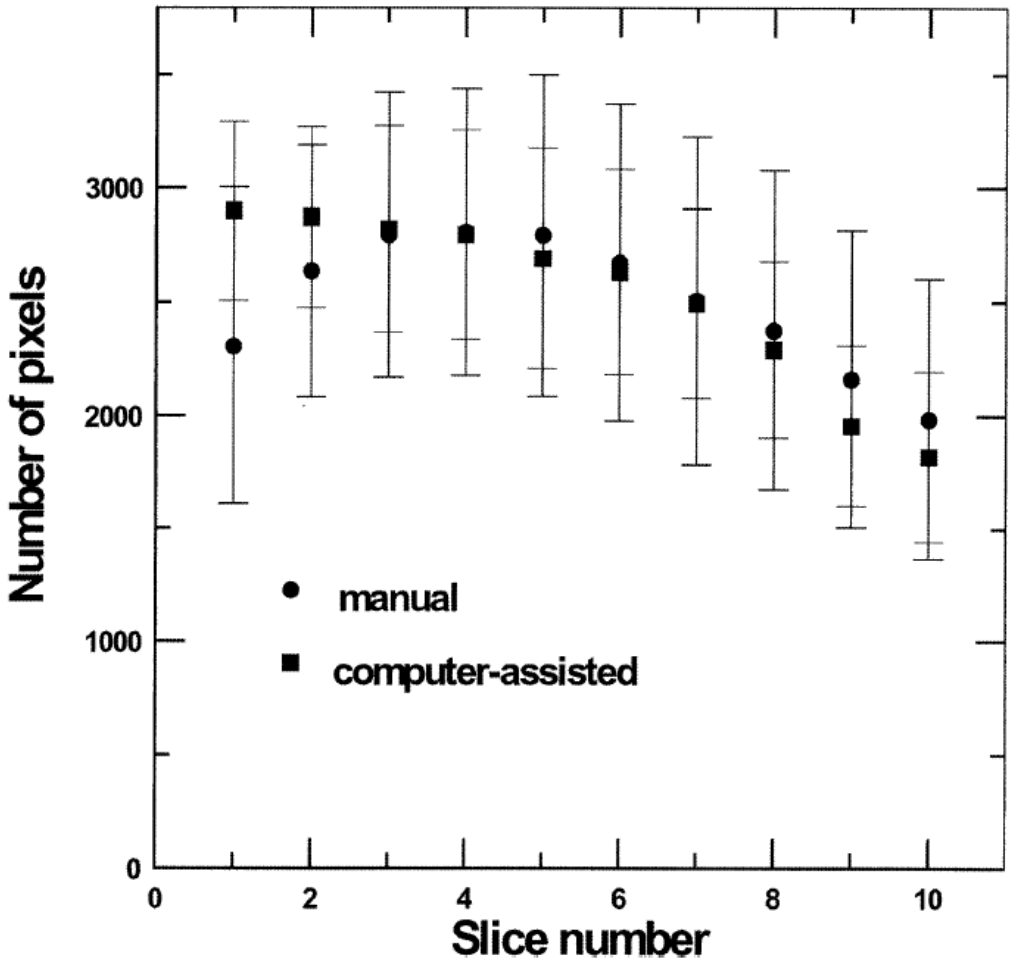


Figura 6.4: Número preomedio de voxeles clasificados como infeccion manualmente (cuadrado) y por el sistema semi-automático cuando la rodaja examinada ha sido usada para entrenar el clasificador. Número promedio de voxeles clasificados como infeccion, sobre cada misma rodaja cuando esa misma se ha usado para entrenar el clasificador del sistema de reconocimiento semi-automatico.

Dia		3	3	3	7	7	7	14	14	14		
Ratón		A	B	C	D	E	F	G	H	I	M	SD
Rod. 1	H	63.1	63.5	88.7	67.9	91.5	64.4	44.1	76.3	80.0	71.1	13.9
	ANN	68.9	82.1	76.6	71.1	89.0	71.6	35.8	71.6	77.6	71.6	14.0
Rod. 2	H	64.2	78.9	65.9	56.9	92.9	50.4	46.3	80.7	51.6	65.3	15.0
	ANN	72.8	86.6	65.4	65.7	88.9	53.0	45.4	79.3	38.4	66.2	16.8

Tabla 6.1: Comparacion del musculo inflamado (en porcentajes) medido por el analisis histopatologico (H) y usando los metodos asistidos por computador (ANN). Sumario de los porcentajes de tejidos inter-lesion inflamatorios (los restantes tejidos se consideran bien necrosis bien acumulacion de esporas) en nueve animales (dos rodajas histologicas por cada uno) inoculados con A. Fumigatus. La fila Dia indica el Número de dias desde la inoculación

la figura 6.4, que incluye todos los datos de los dieciseis ratónes. En esta figura se muestra para cada rodaja la medición automática y la manual obtenidas en promedio. No podemos excluir totalmente la posibilidad de que las diferencias sean debidas también al hecho de que los valores segmentados manualmente para esas rodajas periféricas implican errores experimentales mayores. Quizás ambos factores contribuyen a las diferencias observadas. Las correlaciónes entre las segmentaciones manuales medidas varias veces fueron habitualmente mayores que 0.9 (excepto en uno de los animales estudiados), con significaciones estadísticas menores que 0.01. correlaciónes similares se obtuvieron (excepto para el animal anómalo) entre las medidas manuales y las computerizadas. El número de píxeles dentro de la lesión inflamatoria global determinada para todos los animales tanto manualmente como automáticamente conduce a un coeficiente de correlación de 0.78, con una significación estadística alta (α <0.001). Debería esperarse que se producirá un mejor acuerdo cuando los valores de los volúmenes (la suma de los diez valores de las rodajas) se usen, puesto que el promedio del error de las rodajas individuales se produce de esta manera. Este es en efecto el caso, aunque los valores de los parámetros estadísticos son de un valor ligeramente más alto, la correlación es ahora 0.82.

Un paso más en los estudios de fiabilidad está dado por la comparación del porcentaje de tejido inflamado (o necrosis) identificado por método semiautomático con el identificado por el análisis histopatológico de los tejidos diseccionados (ver tabla 6.1). En este caso, sólo nueve animales y dos secciones de cada uno de ellos

	hipocampo izquierdo		hipocampo derecho		*corpus callosum*	
Paciente	Manual	ANN	Manual	ANN	Manual	ANN
1	2.50	2.55	2.56	2.31	10.44	9.17
2	2.19	2.23	2.35	2.23	9.93	10.67
3	2.34	2.29	2.75	2.71	8.34	8.96
4	1.83	1.94	1.62	1.72	12.16	11.65
Media	2.21	2.25	2.32	2.24	10.2	10.1
SD	0.29	0.25	0.49	0.41	1.6	1.3

Tabla 6.2: Comparacion de las medidas volumetricas (cm3) manuales y automatizadas

fueron seleccionadas, puesto que es muy dificil hacer corresponder las imágenes de resonancia y las muestras histopatológicas, sobre todo por la diferencia en el ángulo de corte. Las medias y desviaciónes estandard dadas en la tabla 6.1 indican que estamos tratando otra vez con medidas paralelas, y por tanto el alto coeficiente de correlación determinado, 0.87 ($p < 0.01$), indica la alta fiabilidad del porcentaje de tejido inflamatorio detectado por el algoritmo.

6.5.2.Datos clínicos

Se presentan en la figura 6.5 los resultados del procedimiento descrito arriba para la segmentación del hipocampo en uno de los cuatro casos clínicos, superimponiendo el área del hipocampo identificada (píxeles blancos) sobre los cortes correspondientes. Sólo una parte pequeña del campo de visión total se ha mostrado para una mejor visualización. Estas imágenes parecen indicar que el método propuesto es capaz de realizar la segmentación de la región objetivo con una precisión que se acerca a la del cerebro y ojo humano. La inspección visual de esta figura permite reconocer que las mayores diferencias entre el método asistido por computador relativo a la segmentación manual se encuentra en las últimas rodajas mostradas, p.ej.: las regiones anatómicas anteriores cercanas al núcleo amigdalítico. La fuente principal de error de sobre segmentación en esta área puede ser el efecto de volúmenes parciales en MRI. Además, el MLP se ha entrenado con las rodajas centrales y tiene problemas para generalizarse a la región periférica, como ya veíamos en el caso animal.

Los volúmenes calculados para cada uno de los cuatro pacientes se dan en la tabla 6.2, donde se aprecia un acuerdo cercano entre la segmentación manual

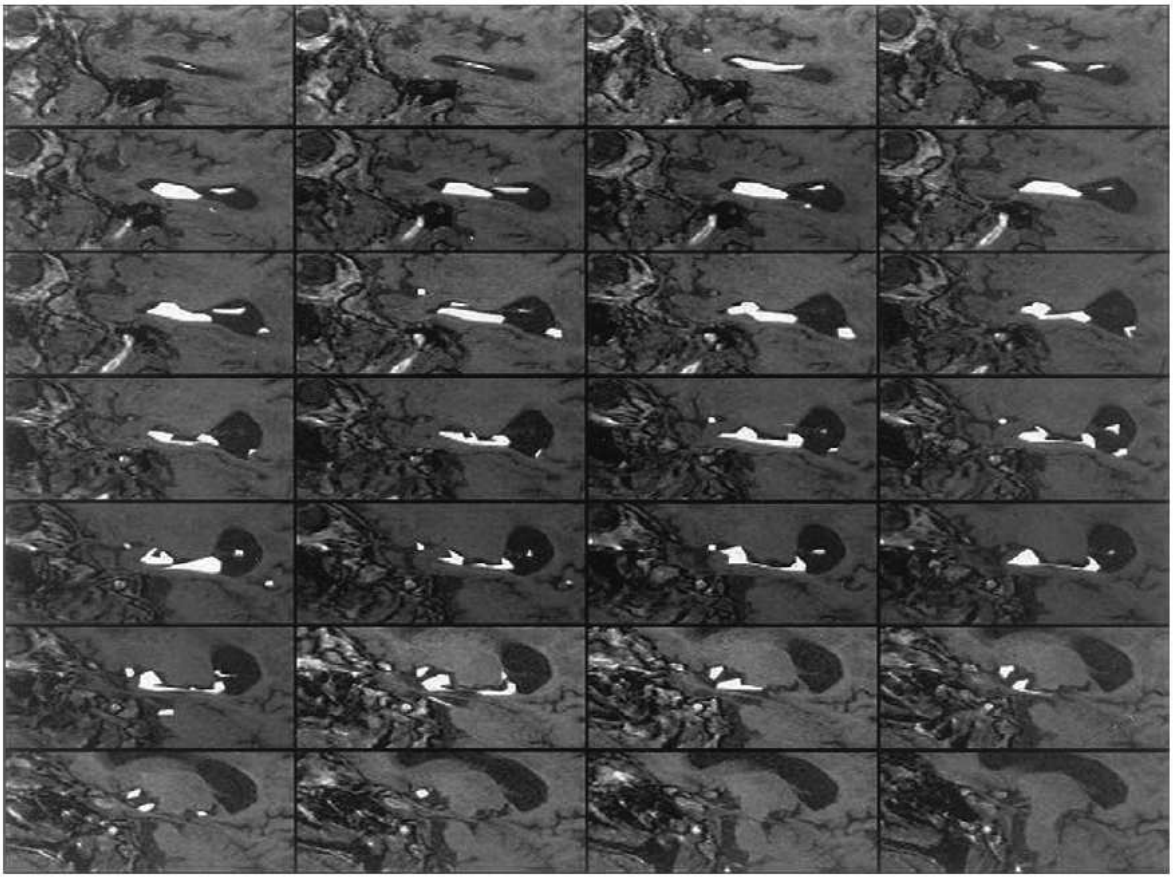


Figura 6.5: Segmentacion de las Imágenes del hipocampo izquierdo, correspondientes a 28 rodajas, usando secuencias de pulsos eco-gradiente T1 estandard. Las areas detectadas por el algoritmo semi-automatico se superponen en cada imagen.

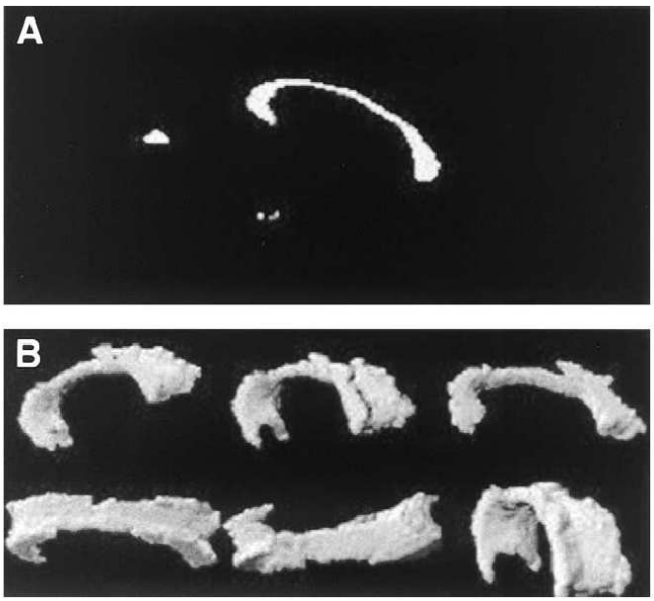


Figura 6.6: A. Area hemi-sagital del Corpus Callosum obtenida por el procedimiento semi-automatico propuesto. B Visualizaciones 3D del Corpus Callosum tras la superesion de los píxeles mal clasificados por un algoritmo de crecimiento de regiones

y los valores del proceso automatizado. Las medidas son paralelas como en el caso del modelo animal. El coeficiente de correlación entre los valores manuales y automatizados es muy alto (r =0.95 para ambos hipocampos), confirmando por tanto la alta fiabilidad observada. Conclusiones similares se obtienen con el ANOVA.

Hemos medido también los volúmenes del the *corpus callosum*. La figura 6.6A presenta una rodaja de los resultados sobre la segmentación del *corpus callosum* por el método asistido por computador. Una detección satisfactoria se observa en la región objetivo, con algunos píxeles espúreos bien separados de la región objetivo. Estas detecciones espúreas corresponden a vóxeles de materia blanca con una distribución espacial y valor cuantitativo T_1 similar al del *corpus callosum*. Estas similaridades hacen que estos vóxeles sean inicialmente marcados por el VQ-BF en la misma clase que el the *corpus callosum* de un total de siete clases de tejidos usados. Como los vóxeles con falsa detección están distantes espacialmente de la región objetivo, pueden ser facilmente eliminados. La figura 6.6B muestras distintas vistas del rendering 3D del corpus callosum tras la supresión de los falsos positivos (que estimamos son un 14%) con la ayuda de métodos de crecimiento de regiones [20]. La forma cóncava característica de esta estructura anatómica, en la que las fibras localizadas en sus bordes se extienden hacia arriba, se puede distinguir facilmente. Los valores cuantitativos dados en la tabla 6.2 indican que los volumenes medidos manualmente son también muy similares a los calculados por la computadora. Las medidas son también paralelas, aunque el coeficiente de correlación es un poco más bajo $r = 0.88$, lo que puede ser esperado de la mayor dificultad experimental. Hemos realizado también un estudio más detallado usando los resultados obtenidos con uno de los paciente, que tiene los menores valores de volumen hipocampal (paciente 4 en la tabla 6.2). Tres rodajas diferentes fueron usadas (cada una por duplicado) para el procedimiento de entrenamiento del MLP. Los volúmenes calculados y medidos manualmente se dan en la tabla 6.3. Un test ANOVA con todos los datos de esta tabla se realizó para determinar si las diferencias observdas eran debidas al hecho de que cada método realmente mide diferente o si son atribuibles a errores aleatorios. Calculamos ambos valores medios al cuadrado (entre e intra) y sus varianzas, dando un valor $F_{(3,8)}$ de 1.73. Podemos concluir (con $p < 0.005$) que no hay diferencias estadísticamente significativas entre los dos métodos al evaluar los volúmenes del hipocampo. Todas las medidas son paralelas, con medias y varianzas estadísticamente iguales de las medidas manuales ($\mu = 1.93$, $\sigma = 0.11$). El coeficiente de correlación entre los datos manuales y semiautomáticos dados en la tabla 6.3 es $r = 0.75$, indicando una alta

Slice	M_1	M_2	C_1	C_2
1	1.83	2.03	1.71	1.81
2	1.90	2.10	1.60	1.60
3	1.94	1.76	1.93	1.82

Tabla 6.3: Volumenes de hipocampo (en cm3) obtenidos manualmente por dos expertos (M1 y M2) y por el metodo automatizado (C1 y C2) usando las tres rodajas centrales del volumen etiquetado manualmente en M1 y M2 para entrenar el MLP

	Modelo animal		Hipocampo
	Areas	Volumenes	Volumenes
S	0.67 (0.14)	0.68 (0.09)	0.67 (0.09)
K_i	0.79 (0.12)	0.80 (0.07)	0.80 (0.07)
TPF	0.76 (0.14)	0.77 (0.09)	0.79 (0.09)
FPF	0.83 (0.17)	0.87 (0.07)	0.90 (0.06)

Tabla 6.4: Indices para un analisis ROC simplificado. (Los valores entre parentesis son las desviaciónes estandard)

fiabilidad.

6.5.3.Resultados basados en el ROC simplificado

Los cuatro índices cuya definición se reprodujo en la sección de métodos y materiales han sido calculados para todas las posibles combinaciones de areas y volúmenes calculados para todos los animales e hipocampos. Estos resultados se sumarizan en la tabla 6.4, mostrando la alta sensitividad y especificidad de las medidas computacionales. Estos índices son casi iguales tanto en la comparación de áreas como de volúmenes en el modelo animal o en el hipocampo. Poseen altos valores y bajas desviaciónes estandard. La similaridad o índice de repetibilidad, S, y el índice Kappa , K_i, va desde 0 a 1, con cero indicando no solapamiento y uno indicando emparejamiento perfecto entre las regiones determinadas por ambos métodos. El índice S es un test más fuerte que K_i, puesto que por ejemplo dos cubos de voxeles de volumen $10 \times 10 \times 10$, desplazados por un voxel a lo largo de la diagonal espacial resulta en un solapamiento de sólo 57% [11]. Nuestros valores

de S y K_i son menores que los otros medidos con algunos fantasmas estandard, mientras que nuestros valores para fracciones positivas ciertas y falsas son más altos que los medidos con los mismos fantasmas [11].

6.6.Discusión y conclusiones

El proceso inflamatorio en uno de los animales experimentales mostraba complejas mezclas de tejidos con una forma pobremente delimitada, difícil de segmentar mediante métodos convencionales [125]. La adecuación del algoritmo propuesto para identificar la lesión total y cuantificar los hallazgos patológicos se muestra claramente en los resultados presentados en la sección anterior. Además, la fiabilidad y validación de la metodología para estos datos se confirma satisfactoriamente por comparación tanto con las imágenes medidas manualmente como con los estudios patológicos subsiguientes, indicando que podrán ser de utilidad en estudios longitudinales futuros de este modelo animal.

La segmentación del hipocampo es de especial interés y dificultad, puesto que esta estructura aparece pobremente definida en las imágenes clínicas rutinarias adquiridas por secuencias 3D rápidas. La delineación de las subregiones en esta estructura se hace usualmente en los planos sagital o coronal. En ambas orientaciones, la separación de la materia gris del gyrus del para-hipocampo y el nucleo amigdalítico no es evidente, pero puede ser muy conveniente, puesto que esta estructura es de importancia para el estudio de la demencia. Específicamente, las rodajas sagitales se usan para el entrenamiento mejorado del MLP en las rodajas centrales (en las que aparece la amígdala) obteniéndose una mejor separación de los tejidos circundantes. Hemos demostrado que nuestra metodología puede ser aplicada satisfactoriamente a este problema.

La estructura del corpus callosum exhibe una forma característicamente elongada en los planos sagitales MRI. Una delineación volumétrica completa del corpus callosum no es inmediata usando procedimientos automáticos o semi-automáticos sin incorporar conocimiento anatómico a priori [11]. La práctica común es medir manualmente su longitud, forma y área en la rodaja central hemi-sagital, a pesar de que otras características morfológicas latentes que han sido descritas para capturar características intrínsecas del *corpus callosum* no son accesibles mediante esas medidas convencionales de tamaño y forma [115]. Esta falta de resultados cuantitativos probablemente se debe a su morfología natural y al inherente pobre contraste entre varias rodajas obtenidas con las secuencias 3D clínicas estándar pesadas en T1, debido a la presencia de tejidos de materia blanca adyacentes. La

atrofia de esta estructura puede estar relacionada con la perdida neuronal *in vivo* en el neocortex. En este sentido, las mediciones 3D de esta estructura pueden proporcionar un criterio diagnóstico efectivo, siendo de importancia particular en pacientes con enfermedades neurodegenerativas. Hemos mostrado la adecuación de nuestra metodología a este difícil problema.

El primer paso de nuestro método es la clasificación VQ-BF de los bloques espaciales de la imagen, usando los elementos representativos de textura generados por un procedimiento de agrupamiento, en imagen MR monoespectrales. Los resultados muestran que la segmentación de MRI se puede realizar en base a la información espacial cuando no hay disponible información multiespectral. El VQ-BF produce la suavización de las imágenes procesadas dependiendo del tamaño del área, preservando las fronteras de las regiones en la imagen, como puede apreciarse en la figura 6.6B. Esta muestra características de preservación de bordes comparable a otras aproximaciones al filtrado de las imágenes, i.e.: el filtrado anisotrópico [43]. El SOM utilizado para la estimación de los representantes de texturas tiene dos propiedades interesantes: (i) es muy robusto frente a condiciones iniciales, y (ii) los representantes resultantes tienden a estar ordenados, debido a las propiedades de preservación topológica del algoritmo. Esta ordenación es importante para los procesos subsiguientes. El papel del VQ-BF es reducir la variabilidad de la intensidad de la señal a lo largo de conjuntos de imágenes [93], proporcionando una clasificación de los voxeles que potencia el rendimiento de los posteriores análisis y procesos de reconocimiento.

El segundo paso de nuestro procedimiento es el entrenamiento supervisado del MLP basado en la selección y delineación manual de las estructuras de interés en una de las rodajas centrales. Este entrenamiento específico del MLP supera el efecto de traslaciones y deformaciones debidas a mal posicionamiento del sujeto y a la evolución de la estructura en el tiempo. Por ejemplo, la localización y forma de la lesión inflamatoria cambia considerablemente de una rodaja a otra a lo largo del estudio serial del modelo animal. El uso de las desviaciónes estándar, desde el centro de masa de la estructura central como una de las entradas a la red podría producir una fuerte sensibilidad a desviaciónes de la forma elipsoidal o circular. Sin embargo, la forma alongada y cóncava del *corpus callosum* se ha detectado con fiabilidad muy alta.

Otros autores han realizado la clasificación y detección del *corpus callosum* y el hipocampo con ANN, pero las imágenes habían sido previamente registradas a un atlas estándar [93]. En ese trabajo, el problema básico de la segmentación robusta contra deformaciones del objeto se evita realizando el registro y la nor-

malización basada en la detección manual de landmarks que permiten el cálculo de la transformada de deformación del objeto. Por otro lado, algunos estudios de segmentación de imágenes MR se restringen a la región objetivo y asumen una buena alineación de la imagen y una intensidad de señal homogénea [45]. Esto alivia el problema de las clases mal balanceadas en el entrenamiento de los clasificadores y evita la necesidad de discriminar clases de tejidos no relacionados con el objetivo de la segmentación. Por el contrario, nuestro método supera esas dificultades extra de tratar el volumen entero de datos, eludiendo la necesidad del recorte manual de la imagen.

Hay tres consideraciones prácticas a tener en cuenta en todo sistema asistido por computador tratando con determinaciones volumétricas. La primera es la reducción en el tiempo máximo de entrenamiento y procesado. La segunda es el empleo de imágenes monoespectrales en el proceso de segmentación, y finalmente, la robustez del procedimiento de segmentación contra variaciones de la anatomía o la forma entre rodajas consecutivas. La primera consideración puede ser críticamente decisiva, particularmente porque estamos considerando usar estos algoritmos con grandes conjuntos de datos, y en ciertas situaciones clínicas. El tiempo teóricamente ahorrado es proporcional al número de rodajas menos una(la utilizada para el entrenamiento) y al tiempo utilizado para delinear la ROI en una rodaja, puesto que durante ese tiempo el radiólogo puede dedicarse a otras tareas. De todas maneras, una comparación en términos de tiempo es muy difícil y requiere la consideración de muchos factores, como la eficiencia del algoritmo y del computador, el número de casos y de imágenes por caso, la estructura anatómica, etc. Independientemente de estas y otras características, se puede esperar un gran ahorro de tiempo incluso para un número pequeño de imágenes. Los tiempos de proceso manual en nuestro método son mucho más cortos que otros métodos semi-automáticos [36]. La segunda consideración es consistente con la práctica diaria de los radiólogos que no disponen habitualmente de datos de alta calidad o mutiespectrales. Hemos mostrado que la segmentación y la determinación volumétrica realizada con datos MRI monoespectrales son fiables en los casos estudiados, de todas maneras, es preferible trabajar con datos multiparametricos para mejorar la discriminación de los tejidos. En nuestro caso, se podría argumentar que al usar la distancia al punto central de la estructura como una entrada al MLP se puede sesgar el algoritmo hacia encontrar regiones con el mismo tamaño que la imagen de entrenamiento. Esto es evidente de los resultados mostrados en la figura 6.4. Sin embargo, está claro que aunque esta opción no es tan flexible como los modelos globales de formas, el método proporciona, como se

puede verificar en la figura, resultados que concuerdan satisfactoriamente con las rodajas centrales usadas para el entrenamiento. En cualquier caso, los modelos deformables son también muy sensibles a las condiciones iniciales y tienen dificultades con los cambios topológicos. Aunque nuestra estrategia no es la solución óptima, no es costosa computacionalmente y ayuda a minimizar la parte manual interactiva y, lo que es más importante, produce resultados fiables en distintas aplicaciones en las que se necesita procesar un gran número de imágenes. En tercer lugar, el procedimiento de segmentación descrito aquí es suficientemente robusto contra las variaciones entre rodajas en anatomía o forma, al menos en los casos estudiados aquí, donde se muestra claramente que la metodología permite la detección de áreas de interés con relativamente pocos voxeles mal clasificados. Contrariamente a otros procedimientos, nuestro método no asegura una superficie globalmente suave entre rodajas. Como vemos en la reconstrucción del *corpus callosum*, la superficie reconstruida contiene inconsistencias que afectan la conformación final pero que son despreciables en la cuantización final del volumen.

En resumen las innovaciones de este trabajo residen en el uso juicioso de una mezcla de métodos supervisados y no supervisados sobre imágenes monoespectrales. En principio la metodología propuesta se puede aplicar a cualquier volumen de datos MRI, así como a imágenes obtenidas con otros técnicas de imagen médica. Más experimentos computacionales y análisis estadísticos son necesarios para la validación global del método propuesto.

7. Determinación del número de clases mediante filtros de Occam

En este capítulo referimos un intento de aplicación de un algoritmo de selección del número de vectores código basado en la idea de que la pérdida producida por un algoritmo de compresión y el ruido pueden cancelarse. En nuestro contexto, el parámetro de control de la compresión es el número de vectores código en el libro de códigos. Ajustar este parámetro para que se produzca la cancelación unicamente del ruido en la imagen es equivalente a tratar de determinar el número de clases de bloques de píxeles que realmente ocurren en la imagen.

En la sección 7.1 se realiza una introducción al problema y a los filtros de Occam. En la sección 7.2 se discute la aplicación de la filosofía de los filtros de Occam al diseño del cuantizador vectorial. En la sección 7.3 se presentan resultados experimentales. Finalmente, en la sección 7.4 se presentan conclusiones para este capítulo.

7.1.Introducción

La aproximación denominada de los filtros de Occam ha sido propuesta por Natarajan en [103], [104], [105]. La formulación original la recogemos en el Apéndice C. Consiste en la aplicación de un algoritmo de compresión con pérdida como un algoritmo de filtrado para aplicaciones de eliminación de ruido aditivo. La aproximación se basa en la cancelación del ruido por la pérdida inducida por el algoritmo en el proceso de compresión/descompresión. Ha sido mostrado en [105] que la aproximación funciona para el ruido aditivo en general si el algoritmo de compresión es *admisible*.

En este capítulo presentamos la aplicación de la aproximación de filtros de Occam a la determinación del tamaño del libro de códigos para el diseño de un cuantizador vectorial (VQ) [42]. Otras aproximaciones a la determinación del número de vectores código encontradas en la literatura consisten en la adición de

términos de complejidad a la función objetivo minimizada para realizar el diseño del libro de códigos [12], o la formulacion de heuristicas para el crecimiento y poda del libro de códigos como en el caso del Growing Neural Gas [37]. La aproximación basada en filtros de Occam da una intuición clara del significado del proceso de selección del tamaño del libro de códigos que se corresponde con una medida de la complejidad de la señal libre de ruido. Esta interpretación es apropiada para los procesos de segmentación no supervisada que realizamos, en particular sobre las imágenes de resonancia magnética. El filtro de Occam implica la estimación de la curva del ratio de compresión frente a la distorsión (ratio-distorsión), lo que lleva consigo la repetida estimación de los libros de código. Para esta tarea hemos empleado, como en los capítulos anteriores, el SOM [80], en un entrenamiento que se realiza en un paso sobre la muestra [51]. Hemos aplicado el método (Occam filter + VQ-BF) a una imagen de resonancia magnética 3D de un embrión humano

7.2.Filtros de Occam y la cuantización vectorial (VQ)

Los filtros de Occam se introdujeron por Natarajan [103], [104] y [105] como un marco general para el diseño de filtros de eliminación de ruido usando algoritmos de compresión con pérdida.

Resumimos las ideas principales asumiendo señales 1D y la métrica Euclidea. Una exposición más detallada se encuentra en el apéndice C. Consideramos una fuente de señal $f = \{f_t; t = 1, 2, ..\}$, y una fuente independiente de ruido aditivo $v = \{v_t; t = 1, 2, ...\}$. La secuencia ruidosa observada es

$$f + v = \{f_t + v_t; t = 1, 2, ..\}. \tag{7.1}$$

Sea C un algoritmo de compresión con pérdida que produce a través de la codificación y decodificación la señal reconstruida $g = \{g_t; t = 1, 2, ..\}$ cuando se aplica a f. La pérdida de C (i.e. el error cuadrático medio) es $\epsilon = \|f - g\| = E\left\{(f_i - g_i)^2\right\}$. El teorema fundamental de convergencia de los filtros de Occam [105] establece que si una señal ruidosa se comprime con un algoritmo de compresión cuya pérdida está ajustada a la magnitud del ruido, el ruido residual en la señal reconstruida es menor que en la secuencia ruidosa original. Esto es, el ruido aditivo y la pérdida se cancelan.

La aplicación práctica de los filtros de Occam depende de la estimación de la magnitud del ruido $\|v\|$. En [104] se propone la siguiente metodología: Estimar la curva de ratio-distorsión y encontrar el punto de inflexión que corresponde al máximo de la segunda derivada. Identificar la coordenada de distorsión del punto

de inflexión con la magnitud del ruido $\|v\|$. Esta identificación se sigue de la idea de que la distorsión como función del ratio de compresión se puede aproximar por dos regiones lineales dependiendo de la relación entre la distorsión de la compresión y la magnitud del ruido.Para $\epsilon \geq \|v\|$ la contribución de la codificación del ruido al ratio de compresión es constante y pequeña. Para $\epsilon < \|v\|$ la contribución de la codificación del ruido al ratio de compresión es grande y domina la codificación de la señal.

Una versión de la definición convencional de la cuantización vectorial [42] para imágnes 3D es como sigue: dado un proceso estocástico cuyo espacio de estados está en el espacio real euclidiano $d \times d \times d$-dimensional, el cuantizador vectorial viene dado por una colección de vectores código $\mathbf{Y} = \{\mathbf{y}_1, .., \mathbf{y}_c\}$ que forman un libro de códigos, donde c es el tamaño del libro de códigos, la operación de codificación es

$$\varepsilon(\mathbf{x}) = \arg\min\{\|\mathbf{x} - \mathbf{y}_i\|; i = 1, .., c\} \tag{7.2}$$

(asumiendo la distancia Euclideana); la operación de decodificación

$$\varepsilon^*(i) = \mathbf{y}_i, \tag{7.3}$$

es la que reconstruye la señal codificada usando el libro de códigos. El diseño del cuantizador vectorial consiste en la estimación del libro de códigos a partir de una muestra $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_n\}$. En nuestro caso hemos utilizado el SOM presentado en el capítulo 2 para calcularlo, dado un tamaño de libro de códigos. Para las aplicaciones de compresión, la imagen se descompone en bloques no solapados y cada bloque se asume como un vector independiente. En VQ-BF los vectores código se consideran referenciados alrededor de su pixel central

$$\mathbf{y}_i = \left(y^i_{l,m,n}; -\frac{d}{2} \leq l, m, n \leq \frac{d}{2} \right). \tag{7.4}$$

Como en el caso de la operación de convolución, la imagen no está descompuesta en bloques. En lugar de ello se considera la ventana $d \times d \times d$ alrededor de cada pixel

$$\mathbf{f}_{i,j,k} = \left[f_{i+l,j+m,k+n}; -\frac{d}{2} \leq l, m, n \leq \frac{d}{2} \right]. \tag{7.5}$$

El proceso de filtrado VQ-BF se define por

$$\widetilde{f}_{i,j,k} = y^{\varepsilon(\mathbf{f}_{ijk})}_{0,0,0}. \tag{7.6}$$

La segmentación VQ-BF es una clasificación que produce

$$\omega_{i,j,k} = \varepsilon\left(\mathbf{f}_{ijk}\right). \tag{7.7}$$

Los vectores código incorporan el modelo probabilístico de los vecindarios de los píxeles. Los procesos de filtrado y clasificación son, respectivamente, las siguientes decisiones MAP

$$p\left(f_{i,j,k} = y^{\omega}_{0,0,0} \,|\mathbf{f}_{ijk}\right) = \delta_{\omega,\varepsilon\left(\mathbf{f}_{ijk}\right)} \tag{7.8}$$

y

$$p\left(\omega_{i,j,k} = \omega \,|\mathbf{f}_{ijk}\right) = \delta_{\omega,\varepsilon\left(\mathbf{f}_{ijk}\right)}. \tag{7.9}$$

La aplicación de los filtros de Occam al diseño de VQ-BF es como sigue:

1. Fijar la definición del espacio de los vectores código.
2. Calcular la curva ratio-distorsión para la compresión VQ usando bloques no solapados extraidos de la imagen
3. Calcular el punto de inflexión en dicha curva, y
4. Aplicar el libro de códigos de tamaño óptimo al procesado de la imagen mediante VQ-BF.

Hay dificultades específicas en la aplicación de la aproximación del filtro de Occam al diseño del cuantizador vectorial. El parámetro de control de este proceso de diseño del VQ es la ratio de compresión, por tanto, el VQ no da una cota de la pérdida si no una estimación de ella. Esta incertidumbre en la pérdida de la compresión implica que la curva de ratio-distorsión será muy ruidosa. La estimación de la magnitud del ruido basada en la detección del punto de inflexión estará también sujeta a incertidumbre. La estimación de la curva de la distorsión frente al ratio de compresión se puede hacer más precisa realizando remuestreos a un alto costo computacional. Además, no existe ningún parámetro (tamaño del libro de códigos, dimensiones de los vectores código) que determine el ratio de compresión. Por tanto, la curva distorsión-ratio no es única. En el proceso de búsqueda hemos calculado distintas curvas ratio-distorsión variando el número de vectores código para cada dimensión de los vectores código. Finalmente, todo el proceso está condicionado a que el algoritmo de compresión VQ sea admisible, lo que recoje la siguiente proposición.

Proposition 1. *Consideremos una señal de entrada cuya distribución de probabilidad puede ser modelada por una mezcla de Gausianas. El VQ dado por el libro de códigos formado con las medias de las Gausianas es un algoritmo de compresión* admisible.

Proof: Si la entrada sigue una distribución que es una mezcla de Gausianas, entonces el error de codificación es una variable aleatoria gausiana de media cero. La esperanza del producto $(f-g)\cdot g$ es trivialmente cero debido a la independencia del error y la señal recuperada tras la compresión y descompresión VQ

$$E\left\{(f-g)\cdot g\right\} = E\left\{(f-g)\right\}\cdot E\left\{g\right\} = 0. \tag{7.10}$$

7.3.Resultados experimentales

La tarea que se trata de realizar es la segmentación no supervisada del volumen 3D de MRI que contiene la imagen de un embrión humano. El volumen consta de 128 rodajas que son imágenes de 128x256 pixels, que pueden representarse por $\left[f_{i,j,k}; i,j = 1,..,128; k = 1,..,256\right]$. Contienen algunos artefactos de iluminación debido al estadio experimental de la bobina de radiofrecuencia. La figure **??**ónfig-embrión muestra algunas de las rodajas tras una reducción de rango de 32 bits/pixel a 8 bits/pixel.

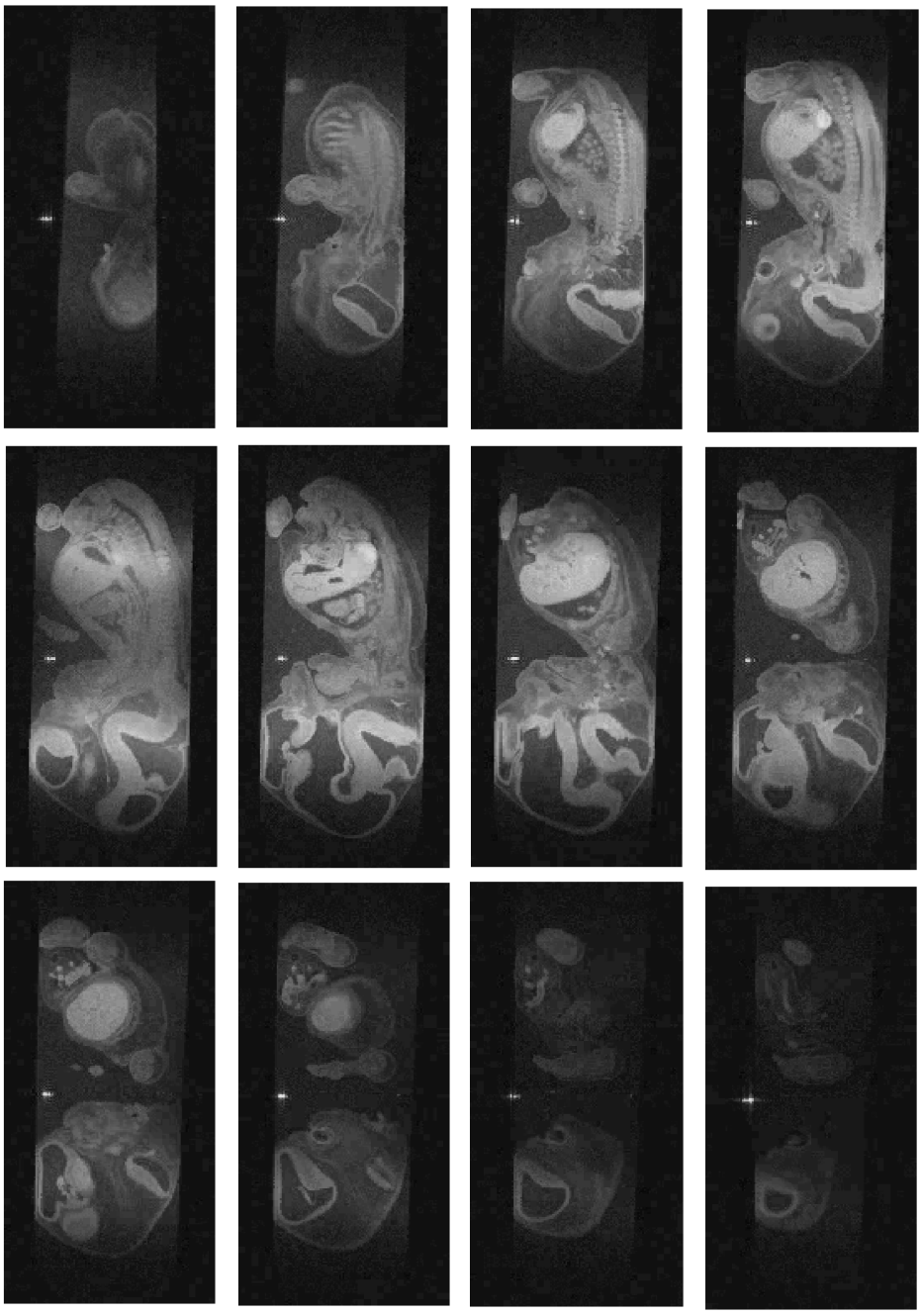
El proceso de segmentación no supervisada consiste en el etiquetado de los píxeles de cada imagen por la aplicación de VQ-BF definido sobre vecindarios 3D, para obtener

$$\left[\omega_{i,j,k}; i,j = 1,..,128; k = 1,..,256\right]. \tag{7.11}$$

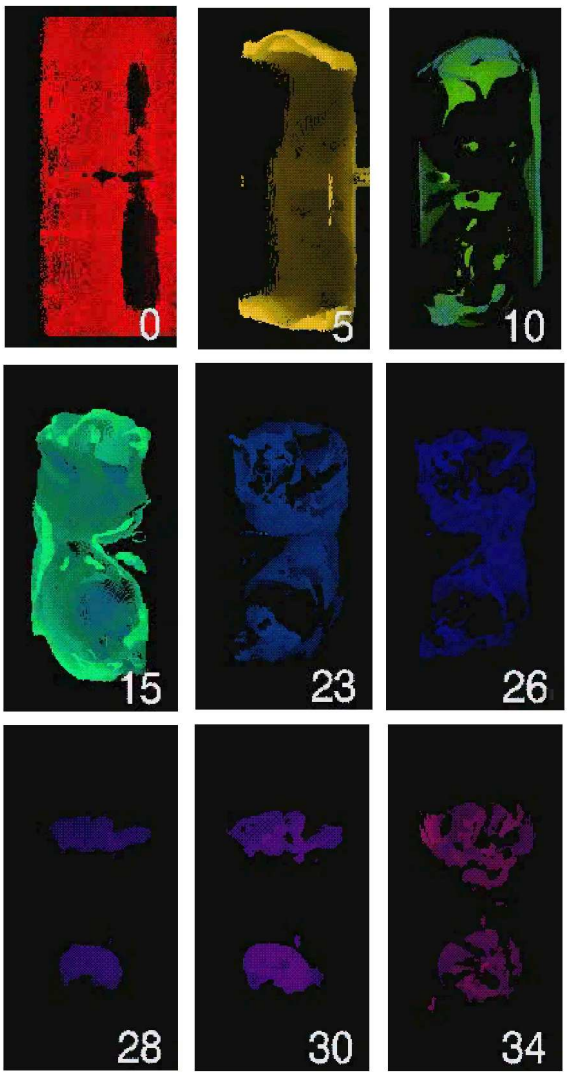
Para visualizar los objetos identificados por cada clase, calculamos las imágenes binarias

$$\left[f^{\omega^*}_{i,j,k}; i,j = 1,..,128; k = 1,..,256\right], \tag{7.12}$$

donde $f^{\omega^*}_{i,j,k} = 1$ si $\omega_{i,j,k} = \omega^*$ y $f^{\omega^*}_{i,j,k} = 0$ en otro caso para $\omega^* = 1,..,c$. Determinamos c por la metodología de los filtros de Occam descrita anteriormente. Cada imagen 3D de una clase se ha renderizado de la siguiente manera: la mitad del volumen se ha considerado rotando alrededor del eje Z , para mostrar la estructura desde diversos puntos de vista. Las vistas en un número de ángulos se han renderizado y recogido en una secuencia. Para cada espacio de vectores código considerado, hemos producido una secuencia con la visualización de todas



Figura˜7.1: Algunas de las rodajas de la imagen 3D de un embrión utilizada en el experimento de determinación del número de clases mediante la aproximación de los filtros de Occam.



Figura˜7.2: Algunas de las vistas de algunas de las clases identificadas con bloques de tamaño 5x5x5

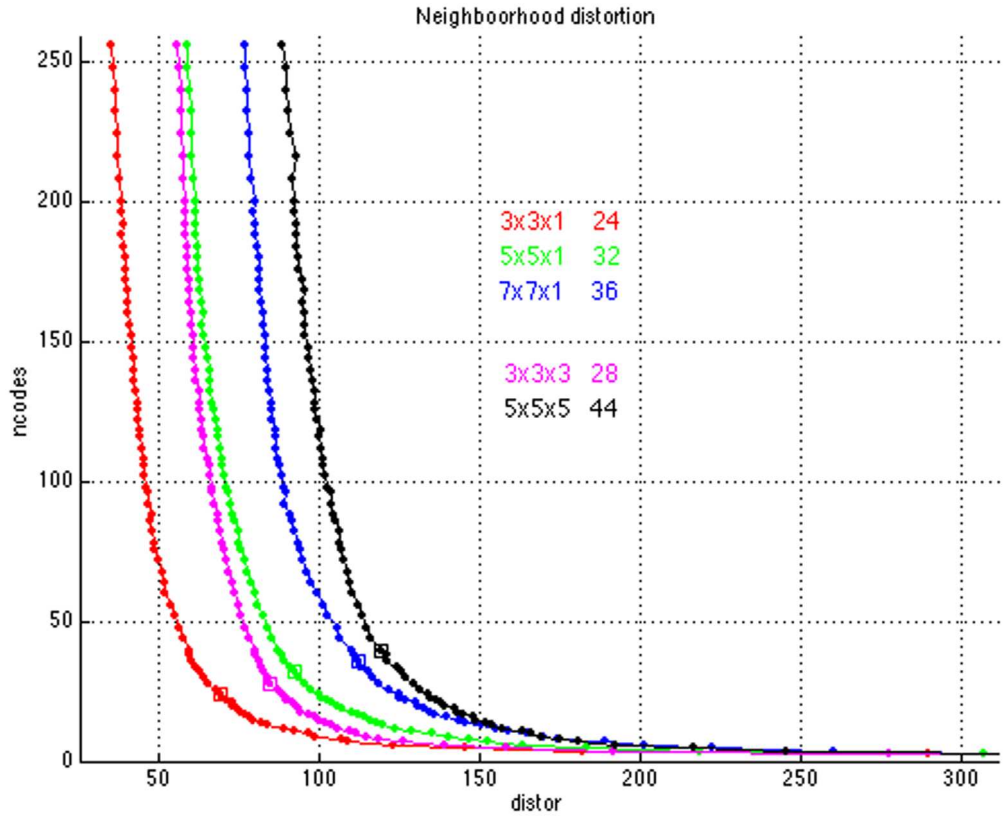


Figura 7.3: Curvas de ratio distorsión calculadas por SOM con un paso sobre la muestra para distintas dimensiones de los vectores codigo.

las clases identificadas por la metodología descrita. Algunas de estas secuencias están disponibles solicitandolas al autor de la memoria.

La figura 7.2 muestra algunas de las vistas obtenidas a partir de la segmentación con bloques 5x5x5 como vectores código. Las orientaciones se han seleccionado arbitrariamente para que resalten las diferencias entre las clases, y cada imagen corresponde a la clase indicada con un número en la propia imagen. El primer comentario se refiere a la continuidad espacial de los objetos obtenidos de esta manera.

La figura 7.3 muestra las curvas de ratio-distorsión y los puntos de inflexión (marcados por pequeños cuadrados) calculados para distintas dimensiones de los

vecindarios. En particular el punto de inflexión para la curva de ratio-distorsión de los vectores código 5x5x5 corresponde a $c = 35$. Volviendo a la figura 7.2, se puede apreciar que las primeras clases corresponden a la solución salina que contiene el embrión. Las clases 3 a 11 parecen identificar las capas frontera entre el embrión y la solución salina. Desde la clase 12 en adelante se identifican volumes interiores del embrión. Algunos de estos volumenes están muy dispersos, otros muestran clases con componentes fuertemente conectados. Las clases 28, 29, 30 identifican dos componentes separados en la imagen que parecen corresponder al cerebro y regiones abdominales. La ordenación de las clases se explica por la propiedad de preservación topológica del SOM. Los primeros vectores código corresponden a las vectores con menor magnitud. La magnitud del vector código aumenta con el índice de la clase, por lo que los tejidos con mayor respuesta de resonancia magnética se identifican con las últimas clases. Podemos apreciar que el método de Occam ha ayudado a mitigar el problema de la falta de balanceado de las clases producido por la solución salina que corresponde a un porcentaje muy elevado de vóxeles en la imagen. Los libros de códigos con tamaño excesivo tienden a asignar muchos vectores código a este fondo salino. El tamaño determinado por el método de los filtros de Occam fuerza su representación parsimoniosa con sólo 3 clases y produce una buena identificación para su eliminación.

7.4. Conclusiones

Hemos trasladado la filosofía de los filtros de Occam al diseño de un libro de códigos óptimo en el sentido del número de vectores código. Esto es, determinamos el número de vectores código ó ptimos como el punto de inflexión de la curva de distorsión versus ratio de compresión. Este proceso implica la realización de un número considerable de cuantizadores vectoriales, para acelerar el proceso se ha realizado sobre una submuestra de la imagen. Una vez determinado el número de vectores código, se realiza la estimación del libro de có digos sobre la imagen completa. Como en capítulos anteriores, el SOM en un paso sobre la muestra es el algoritmo utilizado para la estimación del libro de códigos. El libro de códigos se utiliza en el VQ-BF para el procesado no supervisado de las imágenes de MRI. En este capítulo se muestran resultados sobre una imagen MRI de un embrión humano. Los resultados visuales muestran una gran coherencia espacial de las clases detectadas y un número pequeño de clases espúreas.

Pensamos que todavía se pueden realizar trabajos futuros en esta línea en la mejora de la estimación de la curva ratio-distorsión y de su punto de inflexión.

Un tema de interés es la determinación de la dimensión apropiada de los vectores código, que puede también considerarse en forma jerarquizada.

8. Apéndice A: Introducción al método del gradiente estocástico

Dada la importancia central del gradiente estocástico en el desarrollo de las reglas adaptativas de las redes neuronales y muchos sistemas de computación suave (*Soft Computation*), reproducimos [39] una revisión de su derivación y la prueba de convergencia que clarifica el por qué de las condiciones impuestas a los coeficientes de ganancia en los algoritmos de aprendizaje.

Example 1. *Estimación incremental de la media. Sea una muestra de datos*

$$S = \{\mathbf{x}_1, .., \mathbf{x}_n\},$$

donde los $\mathbf{x}_i$ *son vectores reales, el estimador consistente de la media*

$$\widehat{\mathbf{m}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

puede escribirse de forma incremental como

$$\widehat{\mathbf{m}}_n = \frac{n-1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n-1}\mathbf{x}_i\right] + \frac{1}{n}\mathbf{x}_n = \frac{n-1}{n}\widehat{\mathbf{m}}_{n-1} + \frac{1}{n}\mathbf{x}_n.$$

Esta expresión recuerda fuertemente a las expresiones propuestas para las reglas adaptativas de las redes neuronales competitivas.

8.1. Método de Robins-Monro

Sean z y θ dos variables aleatorias correlaciónadas. El problema a resolver consiste en buscar la raiz de la función de regresión

$$f(\theta) = E[z\,|\theta]$$

Esto es, dado un conjunto de muestras $\{(\theta_i, z_i); i = 1, 2, ...\}$ determinar valor θ_0 para el que $E[z|\theta_0] = 0$.

Si $\widehat{\theta}_N$ es el estimador de θ_0 a la presentación de la muestra (θ_N, z_N) el método de Robins-Monro consiste en calcular

$$\widehat{\theta}_{N+1} = \widehat{\theta}_N - a_N z_N$$

donde $\{a_N\}$ es una secuencia de números positivos que cumplen

1. $\lim_{N\to\infty} a_N = 0$
2. $\sum_{N=1}^{\infty} a_N = \infty$
3. $\sum_{N=1}^{\infty} a_N^2 < \infty$

Puede demostrarse que $\widehat{\theta}_N$ converge a θ_0 en el sentido de la media de los cuadrados (mean square sense)

$$\lim_{N\to\infty} E\left[\widehat{\theta}_N - \theta_0\right] = 0$$

y con probabilidad 1

$$\lim_{N\to\infty} P\left[\widehat{\theta}_N = \theta_0\right] = 1$$

Ejemplos de sequencias de velocidad de aprendizaje

Algunas secuencias que cumplen las condiciones anteriores:

$$a_N = \frac{1}{N}$$

$$a_N = \frac{1}{N^k} \qquad \frac{1}{2} < k \leq 1$$

$$a_N = a_0 \frac{1}{N} \qquad a_0 < 1$$

$$a_N = a_0 \frac{1}{1 + kN} \qquad a_0 < 1; 0 < k < 1$$

$$a_N = a_0 (1 - kN) \qquad a_0 < 1; 0 < k < 1$$

en ocasiones se pueden utilizar expresiones aproximadas aunque no cumplen las condiciones de convergencia

$$a_N = \begin{cases} a_0 \left(1 - \frac{N}{N_{max}}\right) & N < N_{max} \\ 0 & sino \end{cases}$$

Por último, mencionamos el decrecimiento exponencial [100] que es el que utilizamos en muchos de los experimentos computacionales

$$\beta(k) = \beta_{initial} \left(\frac{\beta_{final}}{\beta_{initial}}\right)^{\frac{k}{k\max}}$$

8.2. Convergencia en probabilidad del método de Robins-Monro

Consideremos que z es una v.a. generada a partir de la función de regresión añadiendole ruido

$$z_N = f\left(\widehat{\theta}_N\right) + \gamma_N$$

La aplicación del método de Robins-Monro nos da

$$\widehat{\theta}_{N+1} = \widehat{\theta}_N - a_N z_N = \widehat{\theta}_N - a_N f\left(\widehat{\theta}_N\right) + a_N \gamma_N$$

donde $\gamma_N = z_N - f\left(\widehat{\theta}_N\right)$ es una v.a de media cero (por definición)

$$E\left[\gamma_N \middle| \widehat{\theta}_N\right] = E\left[z_N \middle| \widehat{\theta}_N\right] - f\left(\widehat{\theta}_N\right) = 0$$

y varianza acotada $E\left[\gamma_N^2\right] \leq \sigma^2$.

Además γ_N y $\widehat{\theta}_N$ son estadísticamente independientes (las esperanzas de los productos entre ellos son nulas)

Considerese la distancia entre el estimador y la raiz

$$\left(\widehat{\theta}_{N+1} - \theta_0\right) = \left(\widehat{\theta}_N - \theta_0\right) - a_N f\left(\widehat{\theta}_N\right) + a_N \gamma_N$$

y la esperanza de su cuadrado (el error cuadrático medio), el cual disminuye de la siguiente manera con cada incremento

$$E\left[\left(\widehat{\theta}_{N+1} - \theta_0\right)^2\right] - E\left[\left(\widehat{\theta}_N - \theta_0\right)^2\right] =$$

$$= a_N^2 E\left[f^2\left(\widehat{\theta}_N\right)\right] + a_N^2 E\left[\gamma_N^2\right] - 2a_N E\left[\left(\widehat{\theta}_N - \theta_0\right) f\left(\widehat{\theta}_N\right)\right]$$

Resolviendo recursivamente encontramos

$$E\left[\left(\widehat{\theta}_{N+1} - \theta_0\right)^2\right] - E\left[\left(\widehat{\theta}_1 - \theta_0\right)^2\right] =$$

$$= \sum_{i=1}^{N-1} a_i^2 \left(E\left[f^2\left(\widehat{\theta}_i\right)\right] + E\left[\gamma_i^2\right]\right) - 2\sum_{i=1}^{N-1} a_i E\left[\left(\widehat{\theta}_i - \theta_0\right) f\left(\widehat{\theta}_i\right)\right]$$

Asumiendo que $E\left[f^2\left(\widehat{\theta}_i\right)\right] \leq b$ en la región de interés

$$\underbrace{E\left[\left(\widehat{\theta}_{N+1} - \theta_0\right)^2\right] - E\left[\left(\widehat{\theta}_1 - \theta_0\right)^2\right]}_{\text{(I)}} =$$

$$\underbrace{\underbrace{\sum_{i=1}^{N-1} a_i^2 \left(b + \sigma^2\right)}_{\text{finito por (3)}} - 2\sum_{i=1}^{N-1} a_i \underbrace{E\left[\left(\widehat{\theta}_i - \theta_0\right) f\left(\widehat{\theta}_i\right)\right]}_{\text{(D1)}}}_{\text{(D)}}$$

encontramos que

- (I) está acotado por debajo porque $E\left[\left(\widehat{\theta}_{N+1} - \theta_0\right)^2\right] > 0$ y $E\left[\left(\widehat{\theta}_1 - \theta_0\right)^2\right]$ es finito
- (D1) es siempre positivo dado que $E\left[\left(\theta - \theta_0\right) f\left(\theta\right)\right] \geq 0$ para todo θ, ya que $\left(\theta - \theta_0\right)$ y $f\left(\theta\right)$ siempre tienen el mismo signo

Considerese la proposición

$$\lim_{N\to\infty} E\left[\left(\widehat{\theta}_i - \theta_0\right) f\left(\widehat{\theta}_i\right)\right] = 0$$

si no se cumple entonces (D)$\rightarrow -\infty$, lo cual contradice que (I)$>-\infty$ (está acotado inferiormente)

Por lo tanto la proposición se cumple

Dado que $\left(\theta - \theta_0\right) f\left(\theta\right) \geq 0$ esta proposición equivale a la convergencia en probabilidad del estimador

$$\lim_{N\to\infty} P\left[\widehat{\theta}_N = \theta_0\right] = 1$$

8.3.Aplicación a la minimización: el método del gradiente estocástico

Sea $f(\theta, x)$ una función a minimizar en θ. El mínimo estará en una raiz de $\left\langle \frac{df(\theta,x)}{d\theta} \right\rangle_x$, por lo que el método de Robins-Monro es aplicable si consideramos la derivada de la esperanza en x como la función de regresión.

$$\widehat{\theta}_{N+1} = \widehat{\theta}_N - a_N \frac{df\left(\widehat{\theta}_N, x_N\right)}{d\theta}$$

$$\widehat{\boldsymbol{\theta}}_{N+1} = \widehat{\boldsymbol{\theta}}_N - a_N \qquad \triangle_{\boldsymbol{\theta}} f\left(\widehat{\theta}_N, x_N\right)$$

8.4.Método de la ecuación diferencial ordinaria

El método de la ecuación diferencial ordinaria puede utilizarse para intentar demostrar la convergencia de una regla adaptativa de la que no conocemos explícitamente la función de energía de la que se ha derivado. Sea una regla adaptativa de la forma

$$\widehat{\theta}_{N+1} = \widehat{\theta}_N - a_N \cdot h\left(\widehat{\theta}_N, \mathbf{x}_N\right)$$

La sucesión $\left\{\widehat{\theta}_N; N = 0, 1, ..\right\}$ converge a un mínimo (un subconjunto de estados asintóticamente estable S) de

$$H(\theta) = \lim_{N\to\infty} E_{\mathbf{x}}\left[h(\theta, \mathbf{x})\right]$$

si se cumple que

1. esta esperanza existe para todo posible θ,
2. $P\left[\lim_{N\to\infty} \widehat{\theta}_N = \theta_0\right] = 1$ o bien $P\left[\lim_{N\to\infty} \widehat{\theta}_N \in S\right] = 1$
3. $H(\theta_0) = 0$ y por lo tanto θ_0 es un mínimo de la ecuación diferencial ordinaria

$$\frac{dF}{d\theta} = H(\theta)$$

o bien S es un conjunto asintoticamente estable de dicha ecuación diferencial con dominio de atracción $D\{S\}$

4. x_N está acotada con probabilidad 1

5. $\{a_N\}$ es una secuencia de números positivos que cumplen

   1. $\lim_{N\to\infty} a_N = 0$
   2. $\sum_{N=1}^{\infty} a_N = \infty$
   3. $\exists p$ t.q. $\sum_{N=1}^{\infty} a_N^p < \infty$

Entonces $\widehat{\theta}_N$ entra en un compacto $A \subset D\{S\}$ infinitamente a menudo

Convergencia del Algoritmo competitivo simple

Se puede estudiar, como ejemplo básico, la convergencia del SCL usando el método de la ecuación diferencial ordinaria. Partiendo de la expresión

$$\triangle \mathbf{m}_i(N) = a_N(\mathbf{x}(\mathbf{N}) - \mathbf{m}_i(N))\,\delta_i(\mathbf{x}(\mathbf{N}), \mathbf{m});\ i = 1, .., c$$

Identificamos para cada $\mathbf{m}_i$

$$h\left(\widehat{\theta}_N, \mathbf{x}_N\right) \equiv -(\mathbf{x}(\mathbf{N}) - \mathbf{m}_i(N))\,\delta_i(\mathbf{x}(\mathbf{N}), \mathbf{m})$$

La función de regresión que consideramos es

$$\begin{aligned} H(\mathbf{m}_i) &= \lim_{N\to\infty} E_{\mathbf{x}}[-(\mathbf{x}(\mathbf{N}) - \mathbf{m}_i(N))\,\delta_i(\mathbf{x}(\mathbf{N}), \mathbf{m})] \\ &= E_{\mathbf{x}}[-(\mathbf{x} - \mathbf{m}_i)\,|\mathbf{x} \in \mathcal{R}_i] \end{aligned}$$

El cero de esta función corresponde a

$$E_{\mathbf{x}}[-(\mathbf{x} - \mathbf{m}_i)\,|\mathbf{x} \in \mathcal{R}_i] = 0,$$

por tanto:

$$E_{\mathbf{x}}[\mathbf{x}\,|\mathbf{x} \in \mathcal{R}_i] = \mathbf{m}_i.$$

Lo que significa que el algoritmo competitivo simple converge a los centros de las clases que cuantizan la población.

La función que se minimiza se obtiene integrando la ecuación diferencial

$$\frac{\partial J_i}{\partial \mathbf{m}_i} = H(\mathbf{m}_i) = E_{\mathbf{x}}\left[-(\mathbf{x}-\mathbf{m}_i)\,|\mathbf{x}\in\mathcal{R}_i\right].$$

Se obtiene al realizar la integración:

$$\begin{aligned} J_i &= \int_{\mathbf{m}_i} E_{\mathbf{x}}\left[-(\mathbf{x}-\mathbf{m}_i)\,|\mathbf{x}\in\mathcal{R}_i\right] d\mathbf{m}_i \\ &= E_{\mathbf{x}}\left[\frac{1}{2}\|\mathbf{x}-\mathbf{m}_i\|^2\,|\mathbf{x}\in\mathcal{R}_i\right]. \end{aligned}$$

Si consideramos todos los centros de clase, la función que se minimiza es la distorsión

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} E_{\mathbf{x}}\left[\frac{1}{2}\|\mathbf{x}-\mathbf{m}_i\|^2\,|\mathbf{x}\in\mathcal{R}_i\right]$$

9. Apéndice B: Convergencia del SOM en el caso escalar

El caso escalar es el único en el que hasta el momento se ha probado la convergencia del SOM a una configuración ordenada de los pesos de las unidades neuronales que demuestra el principio de conservación topológica. La prueba original es de Marie Cottrell y la extraemos de [80]. El interés de esta prueba es mostrar el concepto de preservación topológica y las limitaciones formales en el análisis de la convergencia de esta clase de redes neuronales. En su momento supuso el primer trabajo de demostración formal de convergencia topológica. Los trabajos posteriores de Fort y Pages [32], [33], [34] no hacen sino subrayar la dificultad de estudiar la convergencia de este algoritmo en dimensiones superiores y en el caso general. Sólo algunos resultados parciales se han probado, como por ejemplo la no convergencia a una configuración ordenada si se mantiene constante la velociada de aprendizaje. En este contexto, que pretende subrayar la demostración siguiente, se entiende que las afirmaciones sobre convergencia en un solo paso sobre la muestra son difíciles de analizar y probar.

Se considera $x, m_i \in A \subset \mathbb{R}$, los indices definidos en $i, k \in \{1, ., L\}$, la distancia entre indices es la de Manhatann (valores absolutos) y la distancia en el espacio input es tambien en valores absolutos $|x - m_i|$. La función de vecindad (invariante en t) es de la forma

$$v(k,i) = \begin{cases} 1 & k \in \{\max(1, i-1), i, \min(i+1, L)\} \\ 0 & \text{sin } o \end{cases} . \tag{9.1}$$

Proposition 2. *Dados $m_i(0)$ aleatorios, conforme $t \to \infty$, si los x_t son extraidos uniformemente en el espacio input, los representantes $(m_1, .., m_L)$ convergen hacia una secuencia ordenada ascendente o descendente. Una vez que se ordenan los representantes, permanecen ordenados para todo t. Además la función de densidad de m_i se apoxima a alguna función de $p(x)$*

La demostración tiene dos partes

1. La formación de la secuencia ordenada

2. La convergencia a puntos fijos una vez ordenada

Estudiaremos en detalle el proceso de convergencia a la ordenación de los pesos. Se define un índice de desorden

$$D = \sum_{i=2}^{L} |m_i - m_{i-1}| - |m_L - m_1| \,, \tag{9.2}$$

que será $D \geq 0$, y será $D = 0$ cuando los m_i están ordenados

Def 1 Un pliegue de longitud $2k+1$ en el nodo j es una secuencia $(m_{j-k}, .., m_{j+k})$ en las que las secuencias parciales $(m_{j-k}, .., m_j)$ y $(m_j, .., m_{j+k})$ son de tendencia opuesta (una es creciente y la otra decreciente o viceversa)

Def 2 Un valor x cae dentro del pliegue si

$$x \in [m_{j-k}, m_j] \bigcap [m_j, m_{j+k}] \,, \tag{9.3}$$

sino cae fuera del pliegue

Def 3 Un pliegue-borde de longitud k en $j = 1$ $[j = L]$ es una secuencia monótona creciente o decreciente de valores $(m_L, m_1, .., m_k)$ $[(m_1, m_L, .., m_{L-k})]$

Def 4 Un valor x cae fuera de un pliegue-borde en j si hay un pliegue borde en j y los signos de $x - m_1$ y $x - m_L$ son distintos

Def 5 Un pliegue-borde doble de longitud $k+1$ en $j = 2$ $[j = L-1]$ es una secuencia de valores $(m_1, .., m_{k+1}, m_L)$ $[(m_L, .., m_{L-k}, m_1)]$ en la que la secuencia parcial $(m_2, .., m_{k+1})$ $[(m_{L-1}, .., m_{L-k})]$ es monotona creciente o decreciente y además los signos de $m_1 - m_2$ y $m_2 - m_3$ son diferentes $[m_L - m_{L-1}$ y $m_{L-1} - m_{L-2}]$ y además los signos de $m_1 - m_2$ y $m_1 - m_L$ $[m_L - m_{L-1}$ y $m_L - m_1]$ son identicos

Def 6 Un valor x cae fuera de un pligue borde doble en $j = 2$ $[j = L-1]$ si hay un pliegue borde doble en j y los signos de $x - m_2$ y $x - m_L$ $[x - m_{L-1}$ y $x - m_1]$ son diferentes

Theorem 3. *Se cumple*

A Sea i el nodo seleccionado al presentarse un input aleatorio x; si $3 \leq i \leq L-2$ entonces D decrece al aplicar la regla del SOM excepto

A1 D no cambia si $(m_{i-2}, .., m_{i+2})$ forma una secuencia monotona

A2 D crece si x cae fuera de un pliegue de longitud mayor que 5 en i

B Si el nodo seleccionado es $i = 1$ o $i = L$ entonces D decrece excepto

B1 D no cambai si $(m_1, .., m_3, m_L)$ o $(m_L, .., m_{L-2}, m_1)$ forman una secuencia monotona

B2 D crece si x cae fuera de un pliegue borde de longitud mayor que 3 en el nodo i

C Si el nodo seleccionado es $i = 2$ o $i = L - 1$ entonces D decrece excepto

C1 D no cambia si $(m_1, .., m_4, m_L)$ o $(m_L, .., m_{L-3}, m_1)$ forman una secuencia monotona

C2 D crece si x cae fuera de un pliegue borde doble de longitud ≥ 4 en el nodo i

Corollary 4. *Si los $(m_1, .., m_L)$ se ordenan no pueden desordenarse por A1,B1,C1*

Convergencia estocástica: se sigue de la teoria de procesos markovianos

$$\lim_{t \to \infty} P\left[(m_1, .., m_L) \text{ ordenados }\right] = 1 \tag{9.4}$$

La convergencia en distribución se basa en calcular el intervalo S_i de valores input x que afectaran al nodo i, en construir la ecuacion diferencial correspondiente y calcular sus puntos fijos

10. Apéndice C: Filtros de Occam

La idea de utilizar algoritmos de compresión como filtros pasa baja no es nueva, pero Natarajan [103], [104], [105] la ha formulado de una manera más precisa. La idea de utilizar esta aproximación para determinar el número óptimo de clases se basa en la consideración del algoritmo de cuantización vectorial basado en los representantes calculados mediante un algoritmo de agrupamiento, como un proceso de compresión con pérdida. Si tratamos de ajustar el número de clases para conseguir que se cancelen la pérdida y el ruido estaremos determinando el número óptimo de clases para el agrupamiento de los datos. En este apéndice reproducimos los razonamientos originales de la formulación de los filtros de Occam en su versión más general.

Conjetura: Cuando un algoritmo de compresión con pérdida se aplica a una señal con ruido, con la pérdida ajustada a la magnitud del ruido, la pérdida y el ruido tienden a cancelarse en lugar de acumularse.

Introducimos la notación empleada por Natarajan:

- Considera funciones de Baire en el intervalo $[0,1]$ sin pérdida de generalidad
- Para una función f y un número natural n, f_n es la secuencia de muestras de f uniformemente separadas $1/n$. En concreto

$$f_n = \{f(0), f(1/n), f(2/n), ..., f((n-1)/n)\} \tag{10.1}$$

- Una secuencia f_n es un vector en $rset^n$, por tanto, la diferencia entre dos secuencias f_n y g_n es su diferencia vectorial. Una norma en este espacio se denota $\|f_n\|$. La norma l_2 se define $\|f_n\|_2 = \frac{1}{n}\sum_{i=0}^{n-1}(f(i/n))^2$
- Respecto a una norma $\|.\|$ un algoritmo de compresión con pérdida C es un programa que toma como input una secuencia f_n y una tolerancia $\epsilon > 0$ y produce como salida un string binario s que codifica la secuencia g_n cumpliendo $\|f_n - g_n\| < \epsilon$

- Un algoritmo de descompresión D toma como input el string binario $s = C(f_n, \epsilon)$ y produce la secuencia recuperada $g_n = D(s)$
- La pérdida de C (i.e. el error cuadrático medio) es $d = \|f - g\| = E\left\{(f_i - g_i)^2\right\}$.
- v es una v.a. que representa el ruido, v_n es una secuencia de n observaciones independientes de v.
- $\widehat{f_n} = f_n + v_n$ es la secuencia corrompida con el ruido.
- La magnitud del ruido se define como $\|v\| = \lim_{n\to\infty} \|v_n\|$.

Definition 5. *El algoritmo C es* admisible *si el error de reconstrucción $g - f$ y la señal reconstruida g son no correlados, i.e. $(f - g) \cdot g = 0$ donde $f \cdot g = E\{f_i g_i\}$.*

Definition 6. *La función de ratio-distorsión para un codificador C, denotada por $R_C(f, d)$, es el ratio de compresión de la codificación (número promedio de bits por muestra input) para la fuente f como una función de la distorsión d.*

Definition 7. *La función de ratio-distorsión para la fuente f es la mínima posible para cualquier codificador*

$$R(f, d) = \min_C \{R_C(f, d)\}. \tag{10.2}$$

El teorema fundamental de convergencia de los filtros de Occam [105] es el siguiente:

Theorem 5. *Sea g la señal obtenida por la codificación y decodificación de $f + v$ usando un codificador* admisible *C, cuya máxima pérdida está ajustada a $\|v\|$. El ruido residual de la señal reconstruida está acotado por*

$$\frac{\|f - g\|}{\|f\|} \leq \left(2 + \sqrt{2}\right) \sqrt{\frac{R_C(f + v, \|v\|)}{-R'(v, \|v\|)\,\|f\|}} \tag{10.3}$$

donde $R'(v, \|v\|)$ es la derivada izquierda de la función de ratio-distorsión $R(f, d)$ respecto de la distorsión d evaluada en $d = \|v\|$.

La prueba del teorema se puede encontrar en [105]. Este teorema es la formalización de la intuición que conduce a la definición de los filtros de Occam: si una señal ruidosa es codificada por un codificador admisible con la pérdida ajustada a la magnitud del ruido, entonces el ruido y la pérdida tienden a cancelarse, con la extensión de la pérdida en la codificación dependiendo de la incompresibilidad del ruido en relación a la señal.

Algoritmo de Filtrado
input $\widehat{f}_n$
begin Sea $\|v\|$ la magnitud del ruido
Ejecuta $C\left(\widehat{f}_n, \|v\|\right)$
Descomprime para obtener la secuencia filtrada g_n
end
Una explicación intuitiva del proceso podría ser la siguiente:

1. Supuesto que podemos acceder a la fuente de ruido, calculamos la curva de ratio-distorsión para $C\left(v_n, \epsilon\right)$.
    1. Para $\epsilon > \|v\|$ el ratio será muy elevado, porque se puede aproximar el ruido por una constante dentro de esta tolerancia.
    2. Para $\epsilon < \|v\|$ los ratios de compresión serán pequeños y decrecientes al disminuir ϵ
2. Supuesto que podemos acceder a la señal original,calculamos la curva de ratio-distortion para $C\left(f_n, \epsilon\right)$
    1. Para valores altos de ϵ el ratio de compresión sera elevado
    2. Para valores pequeños de ϵ el ratio de compresión sera pequeño
3. Cuando calculamos la curva de ratio-distorsión para la señal corrompida $C\left(\widehat{f}_n, \epsilon\right)$
    1. Para $\epsilon > \|v\|$ la señal domina al ruido y la curva sigue a la de la señal original
    2. Para $\epsilon < \|v\|$ el ruido domina a la señal y la curva sigue a la del ruido.
    3. Para $\epsilon = \|v\|$ la curva muestra un "codo" que puede identificarse con el maximo de la segunda derivada

10.1.La estimación de la magnitud del ruido

La aplicación práctica de los filtros de Occam depende de la estimación de la magnitud del ruido $\|v\|$. En [104] se propone la siguiente metodología: Estimar la curva de ratio-distorsión y encontrar el punto de inflexión que corresponde al máximo de la segunda derivada. Identificar la coordenada de distorsión del punto de inflexión con la magnitud del ruido $\|v\|$. Esta identificación se sigue de la idea de que la función de ratio-distorsión se puede aproximar por dos regiones lineales

$$R_C\left(f+v,d\right) \approx \alpha_1 d; d \geq \|v\| \tag{10.4}$$

y

$$R_C\left(f+v,d\right) \approx \alpha_2 d; d < \|v\| \,. \tag{10.5}$$

Para $d \geq \|v\|$ la contribución de la codificación del ruido al ratio de compresión es constante y pequeña y la pendiente α_1 es pequeña. Para $d < \|v\|$ la contribución de la codificación del ruido al ratio de compresión es grande y domina la codificación de la señal, y la pendiente α_2 es grande.

En situaciones prácticas hemos encontrado que la derivada segunda es muy sensible a la varianza de la estimación de la curva de ratio-distorsión. Hemos estimado las aproximaciones lineales $\alpha_1 d$ y $\alpha_2 d$ a partir de puntos en la curva de ratio-distorsión que se pueden asumir con seguridad en cada una de las regiones consideradas. Seleccionamos el punto de inflexión como el punto en la curva de ratio-distorsión más cercano a la intersección de ambas lineas.

11. Apéndice D: Derivación de FLVQ

En este apéndice damos el desarrollo formal que conduce a la formulación de la regla del FLVQ presentado en el capítulo 2. Proporcionamos este desarrollo porque la expresión que proporcionamos del FLVQ difiere de la encontrada en algunos trabajos de la literatura. Formalmente esta regla consiste en el descenso del gradiente estocástico de la función objetivo J_m que se calcula aplicando la siguiente ecuación en diferencias:

$$\Delta y_i = -\eta_i \frac{\partial J_m}{\partial y_i}. \tag{11.1}$$

Vamos a introducir algunos elmentos de notación para simplificar las ecuaciones. Para un j fijo en $\{1, ..., n\}$, sea

$$\|x_j - y_i\|^2 = \omega_i \tag{11.2}$$

y

$$(u_{ij})^m = \frac{1}{\left[\sum_{s=1}^{c} \left(\frac{\omega_i}{\omega_s}\right)^{\frac{1}{m-1}}\right]^m} = F_i(\omega_1, ..., \omega_c). \tag{11.3}$$

Entonces la expresión de la función objetivo de agrupamiento borroso queda como sigue:

$$\begin{aligned} J_m &= \sum_{j=1}^{n} \left\{(u_{1j})^m \|x_j - y_1\|^2 + ... + (u_{cj})^m \|x_j - y_c\|^2\right\} = \\ &= \sum_{j=1}^{n} \left\{F_1(\omega_1, ..., \omega_c)\,\omega_1 + ... + F_c(\omega_1, ..., \omega_c)\,\omega_c\right\} \end{aligned} \tag{11.4}$$

Descomponemos la derivada de la siguiente forma:

$$\frac{\partial J_m}{\partial y_i} = \frac{\partial J_m}{\partial \omega_i}\frac{\partial \omega_i}{\partial y_i} \tag{11.5}$$

$$= \sum_{j=1}^{n}\left\{\frac{\partial F_1(\omega_1,...,\omega_c)\,\omega_1}{\partial \omega_i} + ... + \frac{\partial F_c(\omega_1,...,\omega_c)\,\omega_c}{\partial \omega_i}\right\}\frac{\partial \omega_i}{\partial y_i}. \tag{11.6}$$

Sea $F_{il}(\omega_1,...,\omega_c)$ la i-esima derivada parcial de $F_l(\omega_1,...,\omega_c)$.

$$F_{il}(\omega_1,...,\omega_c) = \frac{\partial F_l(\omega_1,...,\omega_c)}{\partial \omega_i}. \tag{11.7}$$

Si $l \neq i$ entonces,

$$F_{il}(\omega_1,...,\omega_c) = \frac{\partial\left\{\left[\sum_{s=1}^{c}\left(\frac{\omega_l}{\omega_s}\right)^{\frac{1}{m-1}}\right]^{-m}\right\}}{\partial \omega_i} \tag{11.8}$$

$$= -m\left[\sum_{s=1}^{c}\left(\frac{\omega_l}{\omega_s}\right)^{\frac{1}{m-1}}\right]^{-(m+1)}\frac{\partial\left[\sum_{s=1}^{c}\left(\frac{\omega_l}{\omega_s}\right)^{\frac{1}{m-1}}\right]}{\partial \omega_i} \tag{11.9}$$

$$= -m\left(u_{lj}\right)^{m+1}\frac{1}{m-1}\left(\frac{\omega_l}{\omega_i}\right)^{\frac{1}{m-1}-1}\frac{-\omega_l}{\omega_i^2} \tag{11.10}$$

$$= \frac{m}{m-1}\left(u_{lj}\right)^{m+1}\left(\frac{\omega_l}{\omega_i^m}\right)^{\frac{1}{m-1}}. \tag{11.11}$$

Si $l = i$ entonces,

$$F_{ii}(\omega_1,...,\omega_c) = \frac{\partial\left\{\left[\sum_{s=1}^{c}\left(\frac{\omega_i}{\omega_s}\right)^{\frac{1}{m-1}}\right]^{-m}\right\}}{\partial \omega_i} = \frac{\partial\left\{\left(\omega_i\right)^{\frac{-m}{m-1}}\left[\sum_{s=1}^{c}\left(\omega_s\right)^{\frac{-1}{m-1}}\right]^{-m}\right\}}{\partial \omega_i} \tag{11.12}$$

$$= \frac{\partial\left[\left(\omega_i\right)^{\frac{-m}{m-1}}\right]}{\partial \omega_i}\left[\sum_{s=1}^{c}\left(\omega_s\right)^{\frac{-1}{m-1}}\right]^{-m} + \left(\omega_i\right)^{\frac{-m}{m-1}}\frac{\partial\left\{\left[\sum_{s=1}^{c}\left(\omega_s\right)^{\frac{-1}{m-1}}\right]^{-m}\right\}}{\partial \omega_i} \tag{11.13}$$

$$= \frac{-m}{m-1}\left(\omega_i\right)^{\frac{-m}{m-1}-1}\left[\sum_{s=1}^{c}\left(\omega_s\right)^{\frac{-1}{m-1}}\right]^{-m} + \tag{11.14}$$

$$+\left(\omega_i\right)^{\frac{-m}{m-1}}(-m)\left[\sum_{s=1}^{c}\left(\omega_s\right)^{\frac{-1}{m-1}}\right]^{-(m+1)}\frac{\partial\left[\sum_{s=1}^{c}\left(\omega_s\right)^{\frac{-1}{m-1}}\right]}{\partial\omega_i}$$

$$= \frac{-m}{m-1}\frac{1}{\omega_i}\left(u_{ij}\right)^m - m\left(\omega_i\right)^{\frac{-m}{m-1}}\left[\sum_{s=1}^{c}\left(\omega_s\right)^{\frac{-1}{m-1}}\right]^{-(m+1)}\frac{-1}{m-1}\left(\omega_i\right)^{\frac{-1}{m-1}-1} \tag{11.15}$$

$$= \frac{m}{m-1}\frac{1}{\omega_i}\left[\left(u_{ij}\right)^{m+1} - \left(u_{ij}\right)^m\right]. \tag{11.16}$$

Esto es, en forma resumida

$$F_{il}\left(\omega_1,...,\omega_c\right) = \begin{cases} \frac{m}{m-1}\left(u_{lj}\right)^{m+1}\left(\frac{\omega_l}{\omega_i^m}\right)^{\frac{1}{m-1}} & l \neq i \\ \frac{m}{m-1}\frac{1}{\omega_i}\left[\left(u_{ij}\right)^{m+1} - \left(u_{ij}\right)^m\right] & l = i \end{cases}. \tag{11.17}$$

Por tanto, nos queda la siguiente expresión

$$\frac{\partial J_m}{\partial y_i} = \frac{\partial J_m}{\partial\omega_i}\frac{\partial\omega_i}{\partial y_i} = \tag{11.18}$$

$$= \sum_{j=1}^{n}\left\{\begin{array}{c} F_{i1}\left(\omega_1,...,\omega_c\right)\omega_1 + ... \\ +F_{ii}\left(\omega_1,...,\omega_c\right)\omega_i + F_i\left(\omega_1,...,\omega_c\right) + ... \\ +F_{ic}\left(\omega_1,...,\omega_c\right)\omega_c \end{array}\right\}\frac{\partial\omega_i}{\partial y_i} \tag{11.19}$$

$$= \sum_{j=1}^{n}\left\{\begin{array}{c} \frac{m}{m-1}\left(u_{1j}\right)^{m+1}\left(\frac{\omega_1}{\omega_i^m}\right)^{\frac{1}{m-1}}\omega_1 + ... \\ +\left[\frac{m}{m-1}\frac{1}{\omega_i}\left[\left(u_{ij}\right)^{m+1} - \left(u_{ij}\right)^m\right]\right]\omega_i + \left(u_{ij}\right)^m + ... \\ +\frac{m}{m-1}\left(u_{cj}\right)^{m+1}\left(\frac{\omega_c}{\omega_i^m}\right)^{\frac{1}{m-1}}\omega_c \end{array}\right\}\frac{\partial\omega_i}{\partial y_i} \tag{11.20}$$

$$= \frac{m}{m-1}\sum_{j=1}^{n}\left\{\begin{array}{c} \left(u_{1j}\right)^{m+1}\left(\frac{\omega_1}{\omega_i}\right)^{\frac{m}{m-1}} + ... \\ +\left(u_{ij}\right)^{m+1} - \left(u_{ij}\right)^m + \frac{m-1}{m}\left(u_{ij}\right)^m + ... \\ +\left(u_{cj}\right)^{m+1}\left(\frac{\omega_c}{\omega_i}\right)^{\frac{m}{m-1}} \end{array}\right\}\frac{\partial\omega_i}{\partial y_i} \tag{11.21}$$

$$= \frac{m}{m-1}\sum_{j=1}^{n}\left\{\left[\sum_{l=1}^{c}\left(u_{lj}\right)^{m+1}\left(\frac{\omega_l}{\omega_i}\right)^{\frac{m}{m-1}}\right] - \frac{1}{m}\left(u_{ij}\right)^m\right\}\frac{\partial\omega_i}{\partial y_i}$$

Además tenemos que

$$\frac{\partial \omega_i}{\partial y_i} = -2\left(x_j - y_i\right) \tag{11.22}$$

Por tanto, finalmente obtenemos la expresión

$$\frac{\partial J_m}{\partial y_i} = \frac{-2m}{m-1} \sum_{j=1}^{n} \left\{ \left[\sum_{l=1}^{c} \left(u_{lj}\right)^{m+1} \left(\frac{\omega_l}{\omega_i}\right)^{\frac{m}{m-1}} \right] - \frac{1}{m}\left(u_{ij}\right)^m \right\} \left(x_j - y_i\right) \tag{11.23}$$

Bibliografía

[1] E. Aarts, J. Korst (1989) *Simulated Annealing and Boltzmann Machines* New York: Wiley

[2] S.C. Ahalt, A.K. Khrishnamurty, P. Chen, D.E. Melton (1990) Competitive learning algorithms for vector quantization *Neural Networks* 3:277-290

[3] Atkins M, Mackicwich B. (1998) Fully automated segmentation of the brain in magnetic resonance imaging. *IEEE Tran Med Imaging* 17:98–107.

[4] Baumgartner R, Ryner L, Richter W, Summers R, Jarmasz M, Somorjai R. (2002) Comparison of two exploratory data analysis methods for fMRI: fuzzy clustering vs. principal component analysis. *Magn Reson Imaging* 18:89–94.

[5] T. Berger (1971) *Rate Distortion Theory* Englewood Cliffs, NJ: Prentice Hall

[6] J.C. Bezdek (1981) *Pattern recognition with fuzzy objective function algorithms* New York: Plenum

[7] Bezdek JC, Hall LO, Clarke LP. (1993) Review of MR image segmentation using Statistical Pattern Recognition. *Med. Phys.* 20:1033–48.

[8] J.C. Bezdek, N.R. Pal (1995) Two soft relatives of learning vector quantization *Neural Networks* 8(5) pp 729-743

[9] M.J. Black, Y. Yacoob, S.X. Ju (1997) Recognizing human motion using parameterized models of optical flow, in *Motion Based Recognition* M. Shah, R. Jain (eds) Dordrecht: Kluwer Acad. Pub. pp.245-269

[10] Boone JM.(1993) Neural networks at the crossroads. *Radiology* 189: 357–9.

[11] Bueno G, Musse O, Heitz F, Armspach JP. (2001) Three-dimensional segmentation of anatomical structures in MR images on large data base. *Magn Reson Imaging* 19:73–88.

[12] J. Buhmann, H. Kuhnel (1993) Vector quantization with complexity costs, *IEEE trans. Inf. Theory.* 39:1133-1145

[13] T. Camus (1994) Real-Time Optical Flow, Dept. Comp. Sci. Brown Univ. CS-94-36 1994

[14] T. Camus (1997) Real time quantized optical flow, *Real Time Imaging* 3(2):71-80

[15] K. Chaudhury, R. Mehrotra (1995) A trajectory based computational model for optical flow estimation *IEEE trans Rob. Aut.* 11(5):733-741

[16] O.T.C. Chen , B.J. Sheu, Z Zhang (1994) An adaptive vector quantizer based on the gold-washing method for image compression *IEEE Trans. Circuits Systems for Video Technology* 4(2):143-156

[17] C.K. Chon, C.K. Ma (1994) A fast method of designing better codebooks for image Vector Quantization *IEEE trans. Comm.* 42 (2/3/4):237-242

[18] F.L. Chung, T. Lee (1994) Fuzzy competitive learning *Neural Networks* 7(3):539-551

[19] Clarke LP, Velthuizen RP, Camacho MA, Heine JJ, Vaidyanathan, Hall LO, Thatcher RW, Silbiger ML. (1995) MRI segmentation: methods and applications. *Magn Reson Imaging* 13:343–68.

[20] Cline HE, Lorensen WE, Kikinis R, Jolesz F. (1990) Three-dimensional segmentation of MR images of the head using probability and connectivity. *J Comput Assist Tomog* 14:1037–75.

[21] Cosman P.C, Oehler K.L., Riskin E.A., Gray R.M. (1993) Using Vector Quantization for Image Processing, *IEEE Proceedings* 81(4):1326-1341

[22] M. Cottrell, J.C. Fort, G. Pages (1994) Two or three things that we know about the Kohonen algorithm en *ESANN'94* M. Verleysen (ed) dFacto press, Brussels

[23] E. de Bodt, M. Verleysen, M. Cottrell (1997) Kohonen maps versus vector quantization for data analysis en (ed), *ESANN'97* M. Verleysen dFacto press, Brussels , pp 211-218

[24] Dawant BM, Hanmann SL, Thirion JP, Maes F, Vandermeulen D, Demaerel P. (1999) Automatic 3D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations: Part 1. Methodology and validation on normal subjects. *IEEE Tran Med Imaging* 18:909–26.

[25] Dhawan A.P. (2003) Medical Image Analysis, IEEE Press, New York.

[26] A.H. Dekker (1994) Kohonen neural networks for optimal colour quantization *Network: Comp. Neural Sys.* 5:351-367

[27] D. DeSieno (1988) Adding a conscience to competitive learning *Proc. Int. Conf. Neural Networks* (ICNN'88) vol 1 pp117-124 San Diego

[28] R.O. Duda, P.E: Hart (1974) *Pattern Classification and Scene Analysis*, New York: Wiley

[29] Duda N, Sonka M. (1998) Segmentation and interpretation of MR brain images: an improved active shape model. *IEEE Tran Med Imaging* 17:1049–62.

[30] I. Essa, A. Pentland (1997) Facial expression recognition using image motion, in *Motion Based Recognition* M. Shah, R. Jain (eds) Dordrecht: Kluwer Acad. Pub. pp.271-298

[31] Farrar T.C. (1987) An Introduction To Pulse NMR Spectroscopy, Farragut Press, Chicago.

[32] J.C. Fort, G. Pagès (1995) On the A.S. convergence of the Kohonen algorithm with a general neighboring function *The Annals of Applied Probability* 5(4):1177-1216

[33] J.C. Fort, G. Pagès (1996) Convergence of stochastic algorithms: from Kushner-Clark theorem to the Liapunov functional method *Adv Appl. Prob.* 28:1072-1094

[34] J.C. Fort, G. Pagès (1996) About the Kohonen algorithm: strong or weak Self-organization? *Neural Networks* 9(5):773-785

[35] J.E. Fowler (1996) Adaptive Vector Quantization for the coding of nonstationary sources; PhD Thesis, Ohio University

[36] Freeborough P, Fox N, Kitney R. (1997) Interactive algorithms for the segmentation and quantitation of 3D MRI brain scans. *Comp Meth Programs Biomed* 53:15–25.

[37] B. Fritzke (1994) Growing cell structures: a self-organizing network for unsupervised and supervised learning *Neural Networks* 7:1441-1460

[38] B. Fritzke (1997) A self-organizing network that can follow non-stationary distributions *Proc. ICANN'97*, Berlin: Springer pp.613-618

[39] FukunagaK., (1991) *Introduction to statistical pattern recognition*, Academic Press

[40] J.L. Furlani, L. McMillan, L. Westover (1994) Adaptive colormap selection algorithm for motion sequences, *Proc. Multimedia'94*, San Francisco, Ca, Oct. 1994, pp.341-347

[41] Geman S., Geman D. (1984) Stochastic relaxation, Gibbs Distributions and the Bayesian Restoration of Images *IEEE trans. PAMI* 6:721-741

[42] Gersho A., Gray R.M. (1992) *Vector Quantization and Signal Compression*, New York: Kluwer

[43] Gerig G, Kübler O, Rikinis R, Jolesz FA. (1992) Nonlinear anisotropic filtering o MRI data. *IEEE Trans Med Imaging* 11:221–32.

[44] A. Gianchetti, M. Campani, V. Torre (1998) The use of optical flow for road navigation *IEEE trans Rob.Aut.* 14(1):34-48

[45] Glass JO, Reddick WE. (1998) Hybrid artificial neural network segmentation and classification of dynamic contrast-enhanced MR imaging (DEMRI) of Osteosarcoma. *Magn Reson Imaging* 16:1075–83.

[46] Glass JO, Reddick WE, Goloubeva O, Yo V, Steen RG. (2000) Hybrid artificial neural network segmentation of precise and accurate inversion recovery (PAIR) images from normal human brain. *Magn Reson. Imaging*; 18:1245–53.

[47] Gonzalez, Woods (1989) *Digital image processing*, Academic Press

[48] A.I. Gonzalez, M. Graña, A. d'Anjou, M. Cottrell (1996) On the application of Competitive Neural Networks to Time-varying Clustering problems en F.L Silva, J. Principe, L.B. Almeida (eds) *Spatiotemporal Models in Biological and Artificial Systems* IOS press pp.49-55

[49] A. I. Gonzalez, M. Graña a, A. d'Anjou, F.X. Albizuri, M. Cottrell (1997) Self Organizing Map for Adaptive Non-stationary Clustering: some experimental results on Color Quantization of image sequences en *ESANN'97* M. Verleysen (ed) dFacto press, pp.199-204

[50] A. I. Gonzalez, M. Graña, A. d'Anjou, F.X. Albizuri, M. Cottrell (1997) A sensitivity analysis of the Self Organizing Map as an Adaptive One-pass Non-stationary Clustering algorithm: the case of Color Quantization of image sequences *Neural Processing Letters* , 6:77-89

[51] Gonzalez A. I., Graña M., D´Anjou A., Albizuri F.X., Cottrell M. (1997) A sensitivity analysis of the Self Organizing Map as an Adaptive One-pass Non-stationary Clustering algorithm: the case of Color Quantization of image sequences, *Neural Processing Letter* 6:77-89

[52] González AI, Graña M, Echave I, Ruiz-Cabello J. (1999) Bayesian VQ image filtering design with fast adaptation competitive neural network en *Engineering applications of bio-inspired artificial neural networks* 1999. Vol. II: Springer, Berlin. pp.341-49.

[53] Gonzalez AI, Graña M, Ruiz-Cabello J, d'Anjou A, Albizuri X. (2001) Experimental results of an evolution-based adaptation strategy for VQ Bayesian Filtering. *Inform Sci* 133:249–66.

[54] M. Graña, A.d'Anjou, A.I. Gonzalez, F.X: Albizuri, M. Cottrell (1995) Competitive stochastic neural networks for Vector Quantization of images *Neurocomputing* 7:187-195

[55] Graña M, Echave I, Ruiz-Cabello J, Cortijo M. (2002) Segmentation of infected tissues in IRI using VQ-BF filtering. en *Proc ICSP 2002*, Beijing, China, IEEE Press.

[56] S. Grossberg (1976) 'Adaptive pattern classification and universal recording, I: Parallel development and coding of neural feature detectors' *Biological Cybernetics* 23:121-134

[57] E. Grosso, G. Metta, A. Oddera, G. Sandini (1996) Robust visual sevoing in 3D reaching tasks *IEEE trans. Rob. Aut.* 12(5):732-741

[58] R.M. Haralick, L.G. Shapiro (1993) *Computer and robot vision*, Reading, MA: Addison-Wesley

[59] Harris R.K. (1983) Nuclear Magnetic Resonance Spectroscopy, Pitman, London

[60] S. Haykin (1994) *Neural Networks: A comprehensive foundation* IEEE press, New York: Macmillan Coll. Pub. Co

[61] P. Heckbert (1980) Color image quantization for frame-buffer display *Computer Graphics* 16(3):297-307

[62] J. Hertz, A. Krogh, R.G. Palmer (1991) *Introduction to the theory of Neural Computation* Reading, MA: Addison Wesley

[63] B.K.P. Horn *Robot Vision* MIT Press (1986) (1990)

[64] Jacobs MA, Knight RA, Soltanian-Zadeh H, Zheng ZG, Goussev AV, Peck DJ, Windham JP, Chopp M. (2000) Unsupervised segmentation of multiparameter MRI in experimental cerebral Ischemia with comparison to T2, Diffusion, and ADC MRI parameters and histopatological validation. *J Magn Reson Imaging* 11:425–37.

[65] B. Jahne (1993) *Spatio-Temporal image processing. Theory and Scientific applications.* Berlin:Springer Verlag, LNCS 751

[66] Jain A.K. (1989) *Fundamentals of digital image processing*, Englewood-Clliffs: Prentice-Hall

[67] A. K. Jain, R.C. Dubes (1988) *Algorithms for clustering data* Prentice Hall

[68] R. Jain, R. Kasturi, B.G. Schunck (1995) *Machine Vision* New York: MacGraw Hill

[69] G. Joy, Z. Xiang (1993) Center cut for color image quantization *Visual Computer* 10:62-66

[70] Kahn CE (1996). Decision aids in radiology. *Radiol Clin North Am* 34:607–28.

[71] M.S. Kankahali, B. Mehtre, J.K. Wu (1996) Cluster based color matching for image retrieval *Patt. Recog.* 29(4):701-708

[72] N.B. Karayiannis (1997) A methodology for constructing fuzzy algorithms for learning vector quantization *IEEE trans. Neural Networks* 8(3) pp.505-518

[73] N. B. Karayiannis, A.N. Venetsanopoulos (1993) *Artificial Neural Networks: Learning algorithms, performance evaluation and applications* Norwell, MA: Kluwer Acad. Pub.

[74] N.B. Karayiannis, J.C. Bezdek, N.R. Pal, R.J. Hathaway, P.Pal (1996) Repairs to GLVQ: A new family of competitive learning schemes *IEEE trans. Neural Networks* 7(5):1062-1071

[75] S. Kaski (1997) *Data exploration using Self-Organizing Maps* PhD Thesis, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland

[76] Kelemen A, Székely G, Gerig G. (1999) Elastic model-based segmentation of 3D neuroradiological data sets. *IEEE Tran Med Imaging* 18: 828–39.

[77] R. Klette, K. Schluns, A. Koshan (1999) *Computer vision. Three dimensional data from images* Berlin: Springer-Verlag

[78] T. Kohonen (1982) Clustering, taxonomy, and topological maps of patterns *Proc.. 6th Int. Conf. Pattern Recognition* pp114-128, Munich

[79] T. Kohonen (1989) *Self Organization and Associative memory* Berlin: Springer Verlag

[80] Kohonen T. (1995) *Self Organizing Maps* Berlin: Springer-Verlag

[81] Kohn MI, Tanna NK, Herman GT, Resnick SM, Mozley PD, Gur RE, Alavi (1991) A. Analysis of brain and cerebrospinal fluid volumes with MR imaging. *Radiology* ;178:115–22.

[82] J. Konrad, E. Dubois (1992) Bayesian estimation of motion vector fields *IEEE Patt. Anal. Mach. Int.*, 14(9):910-927

[83] B. Kosko (1991) Stochastic competitive learning *IEEE trans Neural Networks* 2:522-529

[84] B. Kosko (1992) *Neural Networks and fuzzy systems* London: Prentice-Hall Int.

[85] H.K. Kushner, D.S. Clark (1978) *Stochastic approximation for constrained and unconstrained systems* Berlin: Springer Verlag

[86] Kusuzaki K, Shinjo H, Murata H, Takeshita H, Hashiguchi S, Nozaki T, Emoto K, Ashihara T, Hirasawa Y. (2000) Relationship between doxorubicin binding ability and tumor volume decrease after chemotherapy in adult malignant soft tissue tumors in the extremities. *Anticancer Res* 20:3813–6.

[87] R. Lancini, S. Tubaro (1995) Adaptive Vector Quantization for picture coding using neural networks *IEEE trans Comm.* 43(2/3/4):534-544

[88] Liang Z.-P., P.C. Lauterbur (2000), Principles of Magnetic Resonance Imaging. A signal processing perspective. IEEE Press, New York

[89] E. Litmann, H. Ritter (1997) Adaptive color segmentation-A comparison of neural and statistical methods *IEEE trans Neural Net.* 8(1):175-185

[90] T.S. Liu, L.W. Chang (1995) Fast color image quantization with error diffusion and morphological operation *Signal Proc.* 43:293-303

[91] H. Lütkepohl (1993) *Introduction to multiple time series analysis* Springer Verlag, Wien

[92] Mcinerney T, Terzzopoulos D. (1996) Deformable models in medical images analysis: a survey. *Medical Image analysis* 1:91–108.

[93] Magnotta VA, Heckel D, Andreasen NC, Cizadlo T, Corson PW, Ehrhardt JC, Yuh WTC. (1999) Measurement of brain structures with artificial neural networks: two- and three-dimensional applications. *Radiology*; 211:781–90.

[94] C. Malsburg (1973) Self Organization of orientation sensitive cells in the striate cortex *Kybernetic* 14:85-100

[95] J. Mao, A.K. Jain (1996) A Self-Organizing network for hyperellipsoidal clustering *IEEE trans. Neural Networks* 7(1):16-29

[96] T.M. Martinetz, S.G: Berkovich, K.J. Shulten (1993) Neural-gas network for vector quantization and its application to time-series prediction, *IEEE trans. Neural Networks* 4(4):558-569

[97] S. Maybank (1993) *Theory of reconstruction from image motion* Berlin: Springer Verlag

[98] Mcinerney T, Terzzopoulos D. (1996) Deformable models in medical images analysis: a survey. *Medical Image Analysis* 1:91–108.

[99] Metz CE. (1989) Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiology* 24:234–45.

[100] F. M. Mulier, V. S. Cherckassky (1995) Statistical analysis of Self-Organization *Neural Networks* 8(5):717-727

[101] F. M. Mulier, V. S. Cherckassky (1995) Statistical analysis of Self-Organization *Neural Networks* 8(5):717-727

[102] D. Nair, J.K. Aggarwal (1998) Moving obstacle detection from a navigating robot *IEEE trans Rob. Aut.* 14(3):404-416

[103] Natarajan B.K. (1993). Filtering Random Noise via Data Compression, *Proc. IEEE Data Compression Conference*, Snowbird, Utah, pp. 60-69.

[104] B.K. Natarajan (1995) Filtering Random Noise from Deterministic Signals via Data Compression, *IEEE Trans. on Signal Processing*, 43(11):2595-2605.

[105] B. Natarajan, K. Konstantinides, C. Herley (1998) Occam filters for stochastic sources with application to digital images, *IEEE trans. Signal Proc.* 46(5):1434-1438

[106] W.J. Niessen, J.S. Duncan, LMJ Florack, BM ter Haar Romeny, MA Viergever (1995) Spatiotemporal operators and optic flow *Proc. Workshop on physics-based modelling in computer vision,* IEEE Computer Soc. Press. pp-78-84

[107] M.T. Orchard, C.A. Bouman (1991) Color quantization of images *IEEE trans. Signal Proc.* 39(12):2677-2690

[108] N.R. Pal, J.C. Bezdek, R.J. Hathaway (1996) Sequential Competitive Learning and the Fuzzy c-Means Clustering Algorithms *Neural Networks* 9(5) pp. 787-796

[109] T. Pappas (1992) An adaptive clustering algorithm for image segmentation *IEEE trans. Signal Proc.* 40(4):901-913

[110] N. P. Papanikolopoulos, P.K. Koshla, T. Kanade (1993) Visual tracking of a moving target by a camera mounted on a robot: A combination of control and vision *IEEE trans. Rob. Aut.* 9(1):14-34

[111] D.C. Park, I. Dagger (1994) Gradient based fuzzy c-means (GBFCM) algorithm *Proc ICCNN'94* vol 3 pp.1626-1631

[112] Parker J.A. (1990), Image reconstruction in radiology, CRC Press, Boca ratón, CA

[113] Patterson DW. *Artificial neural networks. Theory and applications.* Prentice Hall

[114] R. Perez de Alejo, J. Ruiz-Cabello, M. Cortijo, I. Echave, J. Regadera, J. Arrazola, P. Avilés, P. Barreiro, D. Gargallo, M. Graña, (2003), Computer-assisted enhanced volumetric segmentation magnetic resonance imaging data using a mixture of artificial neural networks, *Magnetic Resonance Imaging* 21(8):901-912

[115] Peterson BS, Feineigle PA, Staib LH, Gore JC. (2001) Automated measurement of latent morphological features in the human corpus callosum. *Human Brain Mapp* 12:232–45.

[116] W.K. Pratt (1991) *Digital image processing* 2nd ed. Wiley Interscience

[117] H.S. Rhee, K.W. Oh (1996) A design and analysis of objective function-based unsupervised neural networks for fuzzy clustering *Neural Processing Letters* 4(2):83-95

[118] H. Ritter, T. Martinetz, K. Schulten (1992) *Neural computation and self-organizing maps: an introduction* Reading, MA: Addison-Wesley

[119] Rizzo G, Scifo P, Gilardi M, Bettinardi V, Grassi F, Cerutti S, Fazio F. (1997) Matching a computerized brain atlas to multimodal medical images. *Neuroimage* 6:59–69.

[120] K Rose, E. Gurewitz, G.C. Fox (1990) A deterministic annealing approach to clustering *Pattern Recognition Letters* 11:589-594

[121] K Rose, E. Gurewitz, G.C. Fox (1990) Statistical mechanics and phase transition in clustering *Physical Review Letters* 65:945-948

[122] K. Rose, E. Gurewitz, G.C. Fox (1992) Vector Quantization by Deterministic Annealing *IEEE trans. Inf. Theory* 38(4):1249-1257

[123] K. Rose, E. Gurewitz, G.C. Fox (1993) Constrained Clustering as an optimization method *IEEE trans. Patt. Anal. Mach. Int.* 15(8):785-794

[124] Rudin M, Beckmann N, Porszasz R, Reese T, Bochelen D, Sauter A. (1999) In vivo magnetic resonance methods in pharmaceutical research: current status and perspectives. *NMR Biomed* 12:69–97.

[125] Ruiz-Cabello J, Regadera J, Santisteban C, Graña M, Pérez de Alejo R, Echave I, Avilés P, Rodríguez I, Santos I, Gargallo D, Cortijo M. (2002) Monitoring acute inflammatory processes in the mouse muscle by MR imaging and spectroscopy: a comparison with pathological results. *NMR Biomed* 15:204–14.

[126] Rusinek H, de Leon MJ, George AE, Stylopoulos LA, Chandra R, Smith G, Rand T, Mourino M, Kowalski H. (1991) Alzheimer's disease: measuring loss of cerebral gray matter with MR images. *Radiology* 178:109–14.

[127] Saeed N. (1998) Magnetic resonance image segmentation using pattern recognition, and applied to image registration and quantitation. *NMR Biomed* 11:157–67.

[128] Schempp W.J. (1998), Magnetic Resonance Imaging, John Wiley & Sons, New York, NY

[129] Shen D, Moffat S, Renick SM, Davatzikos Ch. (2002) Measuring size and shape of the hippocampus in MR images using a deformable shape model. *Neuroimages* 15:422–34.

[130] Soltanian-Zadeh H, Peck DJ, Windham JP, Mikkelsen T. (1998) Brain tumor segmentation and characterization by pattern analysis of multispectral NMR images. *NMR Biomed* 11:201-8.

[131] G.P. Stein, A, Shasua (2000) Model based brightness constraints: on direct estimation of structure and motion *IEEE trans. Patt. Anal. Mach. Int.* 22(9):992

[132] C. Stiller, J. Konrad (1999) Estimating motion in image sequences *IEEE Signal Processing Magazine*, 16(4):70-91

[133] J. Sum, C. Leung, L. Chan, L. Xu (1997) Yet another algorithm which can generate topography map *IEEE trans Neural Networks* 8(5):1204-1207

[134] M. Tistarelli (1996) Multiple constraints to compute optical flow *IEEE Trans. Patt. Anal. Mach. Int.* 18(12):1243-1250

[135] J.T. Tou, R.C. Gonzalez (1974) *Pattern recognition principles* Reading, Ma: Addison-Wesley

[136] E.C:K. Tsao, J.C. Bezdek, N.R. Pal (1994) Fuzzy Kohonen clustering networks *Patt. Recog.* 27(5) pp757-764

[137] T. Uchiyama, M.A. Arbib (1994) Color image segmentation using competitive learning *IEEE trans. Patt. Anal. Mach. Int.* 16(12):1197-1206

[138] J. Vaisey, A. Gersho (1988) Simulated Annealing and codebook design *Proc. ICASP 1988* pp.1176-1179

[139] M.M.van Hulle, D. Martinez (1993) On an unsupervised learning rule for scalar quantization following the maximum entropy principle *Neural Computation* 5:939-953

[140] M.M.van Hulle, D. Martinez (1994) On a novel unsupervised competitive learning algorithm for scalar quantization *IEEE trans. Neural Networks* 5: 498-501

[141] M.M. van Hulle (1997) Topology preserving Map formation achieved with a purely local unsupervised competitive learning rule *Neural Networks* 10(3):431-446

[142] D. Vernon, M. Tistarelli (1990) Using camera motion to estimate range for robotic parts manipulation *IEEE trans. Robot. Autom.* 6(5):509-521

[143] V.V. Vinod, S. Chaudhury, J. Mukherjee, S. Ghose (1994) A connectionist approach for clustering with applications in image analysis *IEEE trans. Sys. Man Cyb.* 24(3):365-383

[144] A.G. Voloboj (1993) The method of dynamic palette construction in realistic visualization systems *Comp. Graph. Forum* 12:289-296

[145] B. Widrow, S.D. Stearns (1985) *Adaptive signal processing* Prentice-Hall

[146] L. Wilxson (2000) Detecting salient motion by accumulating directionally consistent flow *IEEE trans. Patt. Anal. Mach. Int.* 22(8):774

[147] Winkler G. (1995) *Image analysis, random fields and dynamic Monte Carlo methods* Berlin: Springer Verlag

[148] Z. Xiang (1997) Color image quantization by minimizing the maximum intercluster distance *ACM trans. Graphics* 16(3) pp.260-276

[149] Z. Xiang, G. Joy (1994) Color image quantization by agglomerative clustering, *IEEE Comp. Graph. Appl.* pp.44-48

[150] X. Wu (1991) Efficient Statistical Computations for Optimal Color Quantization en *Graphics Gems II*, J. Arvo (ed) Academic Press Professional pp.126-133

[151] Wu Y, Giger ML, Doi K, Vyborni CJ, Schmidt RA, Metz CE. (1993) Artificial Neural Networks in mammography: applications to decision making in the diagnosis of breast cancer. *Radiology* 187:81–7.

[152] Y. Yacoob, L.S. Davis (1997) Estimating image motion using temporal multi-scale models of flow and acceleration in *Motion Based Recognition* M. Shah, R. Jain (eds) Dordrecht: Kluwer Acad. Pub. pp.17-37

[153] E. Yair, K. Zeger, A. Gersho (1992) Competitive learning and soft competition for vector quantizer design *IEEE trans Sign. Proc.* 40(2):294-309

[154] Z. Zhang, V.K. Wei (1996) An on-line universal lossy data compression algorithm via continuous codebook refinement- Part I: Basic results *IEEE trans. Inf. Theory* 42(3):803-

[155] K. Zeger, J. Vaisey, A. Gersho (1992) Globally optimal vector quantizer design by stochastic relaxation *IEEE trans. Signal Proc.* 40(2):310-322

[156] Zeng X, Staib L, Schultz R, Duncan J. (1999) Segmentation and measurement of the cortex from 3D MR images using coupled-surfaces propagation. *IEEE Trans Med Imaging* 18:927–37.

[157] Zijdenbos A, Dawant B, Margolin R, Palmer A. (1994) Morphometric analisis of white matter lesions in MR images: method and validation. *IEEE Tran Med Imaging* 13:716–24.