Vniversitat de València

Departament d'Òptica

Eliminación de Redundancia en el Sistema Visual Humano: Nueva formulación y Aplicaciones a la Codificación de Imágenes y Vídeo

Tesis Doctoral presentada por:

Jesús Malo López

Mayo 1999

JOSE MARÍA ARTIGAS VERDE Profesor Titular de Universidad del Departament d'Òptica de la Universitat de València y
FRANCESC FERRI RABASA, Profesor Titular de Universidad del Departament d'Informàtica i Electrònica de la Universitat de València

CERTIFICAN:

Que la presente memoria, *Eliminación de Redundancia en el Sistema Visual Humano: Nueva Formulación y Contribuciones a la Codificación de Imágenes y Vídeo*, resume el trabajo de investigación realizada bajo su dirección en la Facultad de Física de la Universitat de València por Jesús Malo López, y constituye su Tesis para optar al grado de Doctor en Ciencias Físicas.

Y para que así conste presentan la referida memoria, firmando el presente certificado en Valencia, 10 de Mayo de 1999.

J.M. Artigas Verde F.J. Ferri Rabasa

Índice General

Capítulo 1

Introducción

El problema genérico de la visión entendida como la elaboración de conceptos sobre el entorno a partir de una señal espacio-temporal con características espectrales concretas, involucra una serie de problemas computacionales independientes del tipo de sistema que los resuelva.

Algunos de estos problemas intrínsecos a diferentes niveles de abstracción, han sido formulados por diversas comunidades científicas ajenas, en principio, al estudio del Sistema Visual Humano (SVH), como las dedicadas al procesado de señales, el reconocimiento de patrones o la visión por computador.

Es evidente que el SVH se encuentra (y resuelve satisfactoriamente) estos mismos problemas. Por lo tanto, por una parte, su funcionamiento debe poderse expresar, y explicar, en los mismos términos en los que estas otras disciplinas desarrollan los sistemas que proponen; y por otra parte, los modelos perceptuales desarrollados de esta forma constituirán soluciones particulares (de eventual interés práctico) a esos problemas genéricos.

Estas consideraciones generales, ampliamente aceptadas [1–5], son particularmente apropiadas, y tienen un interés específico, en el caso del estudio de la detección y la discriminación de patrones espacio-temporales simples.

A este (bajo) nivel de abstracción, aparecen problemas generales de gran interés en reconocimiento de patrones y en visión por computador como el de la selección de características cualitativamente significativas o selección del espacio de representación [4, 6, 7], o el problema de la identificación de agrupamientos en esos espacios de representación [6, 7], estrechamente ligado con las características geométricas y topológicas de dichos dominios.

Las técnicas de codificación de imágenes y secuencias, tradicionalmente ligadas al ámbito del procesado de señal y la teoría de la información, se han fundamentado en el análisis de las señales a este mismo nivel de abstracción. De esta forma, también se ha abordado desde este punto de vista el problema de la búsqueda de una transformada o banco de filtros para la representación de la señal [3, 8] y el problema de su simplificación en este nuevo dominio [9], que también requiere de una adecuada prescripción para la medida de distancias.

Así mismo, en el intento de eliminar eficientemente los datos temporalmente redundantes según la clásica idea de la codificación predictiva, se han desarrollado métodos de estimación de movimiento [10–12].

Los modelos de la percepción humana a este nivel proponen soluciones concretas y heurísticas alternativas a los problemas de representación de texturas [13–28] o del movimiento [29–33]. Así mismo, también plantean métodos para evaluar diferencias perceptuales entre entre patrones en las citadas representaciones [19, 34–39].

La simbiosis entre todos estos campos de investigación resulta evidente al ver como, simultaneamente al desarrollo formal de la teoría de las representaciones sub-banda y multirresolución en procesado de señal [3, 40–43], se han planteado representaciones mediante un conjunto de filtros pasa-banda en el SVH [17, 18, 44–47]; y cómo las técnicas de codificación de nueva generación apuntan, por un lado, a la utilización de representaciones de la señal de más alto nivel, tradicionalmente restringidas al ámbito de la visión por ordenador [48–50], y por otro lado, a la consideración de las propiedades de la percepción humana de forma más fundamental en el diseño de los codificadores de señales visuales [51, 52].

En este contexto, resulta interesante profundizar en el estudio de la visión humana a este nivel proponiendo modelos *bien formulados*, es decir, que acepten señales de las dimensiones adecuadas, en los dominios adecuados e interpretables según los parámetros utilizados en reconocimiento de formas o en teoría de la información. El interés en este tipo de formulación se basa en dos argumentos, uno de carácter aplicado y otro de tipo fundamental. Por un lado, los modelos así formulados podrán incluirse, o servir de base, a aplicaciones de comunicaciones en banda estrecha de gran interés en la actualidad [53]. Por otro lado, más importante, este tipo de formulación permite trabajar con escenas reales en condiciones no controladas, de modo que es posible valorar si los refinamientos introducidos a partir de experiencias de laboratorio con estímulos simples tienen relevancia en condiciones naturales de visión [54]. Este tipo de formulación permite decidir si ciertos comportamientos son sólo unos *efectos secundarios* de una determinada implementación biológica de un proceso, o si, por el contrario, tienen una relevancia general porque representan una solución diferente o unas restricciones útiles que mejoran los resultados de algún algoritmo genérico sin inspiración biológica.

1.1 Problemas básicos en compresión de señales visuales: movimiento, representación y cuantización

Los métodos de codificación tienen por objeto expresar la señal de forma no redundante. En las secuencias de imágenes naturales destinadas a ser analizadas por personas existen dos tipos de redundancia: redundancia objetiva y redundancia subjetiva. La redundancia *objetiva* está relacionada con las dependencias de las muestras de una secuencia natural con las muestras de su entorno. En una secuencia con objetos estructurados y movimiento coherente no todas las muestras son necesarias porque parte de ellas pueden predecirse a partir del entorno. La redundancia *subjetiva* está relacionada con los mecanismos de la percepción humana. Ciertos datos de una secuencia natural son irrelevantes para un obsevador humano, porque son eliminados en las primeras etapas de la percepción. Por ello, una expresión de la señal destinada a un observador humano puede prescindir de los datos que sean irrelavantes para el observador aunque no sean predecibles por el contexto.

Básicamente existen dos formas de eliminar la redundancia de una señal:

- Los procesos de compresión sin pérdidas (sin distorsión) basados en encontrar un código binario cuya longitud promedio sea igual a la entropía (de orden n) de la señal [10, 55].

- Los procesos de compresión con pérdidas (con distorsión) basados en la cuantización de una representación de la señal.

Los procesos de compresión sin pérdidas consiguen unas tasas de compresión limitadas. Esto es así porque suelen asumir un comportamiento estadístico demasiado simple que no explota ninguna de las redundancias citadas. Un código basado en la entropía de orden cero, que asume que las muestras son independientes entre si, suele ser muy inadecuado en el caso de la representación espacio-temporal de secuencias naturales.

La cuantización consiste en asignar un representante discreto a cada uno de los puntos de un dominio continuo. La cuantización de una señal implica una reducción de sus grados de libertad y una reducción intrínseca de su entropía. Por contra, la cuantización introduce una distorsión irreversible en la señal. En los algoritmos de compresión con pérdidas (con cuantización) hay que alcanzar una relación de compromiso entre la tasa de compresión deseada y la distorsión tolerable [9, 12, 56, 57].

Los métodos más extendidos de codificación de video utilizan dos técnicas básicas para resolver el problema de la eliminación de la redundancia de la señal: la compensación del movimiento y la cuantización de la señal de error residual en un determinado dominio transformado [10–12, 58, 59].

Compensación de movimiento

La compensación del movimiento consiste en la predicción de muestras futuras a partir de muestras previas y de una cierta información del movimiento de la secuencia. En un caso ideal la información de movimiento y la secuencia previa bastarían para reconstruir exactamente la señal posterior. Con las técnicas actuales de representación del movimiento no es posible esta reconstrucción ideal, de manera que para obtener una reconstrucción perfecta es necesario además una señal residual para compensar los errores de predicción.

La compensación del movimiento implica, pues, representar la fuente de información original (la secuencia) mediante dos nuevas fuentes, la información de movimiento y los errores de predicción. La ventaja de esta representación es que se elimina parte de la redundancia temporal objetiva de forma que la entropía (de orden cero) de las nuevas señales es menor.

Evidentemente en este esquema existe una relación entre el esfuerzo dedicado a la estimación del movimiento y a codificación de la señal residual. Por un lado, una descripción del movimiento más completa (más voluminosa) permitirá una mejor predicción y originará una señal de error de menor entropía. Por otro lado, una mejor codificación de la señal de error (que explote las redundancias no eliminadas por la compensación de movimiento) dará un buen resultado incluso con una información de movimiento simple (de poco volumen y poco predictiva). Esta relación de compromiso es uno de los problemas básicos más importantes en los esquemas actuales de compresión de video [12, 59, 60].

Transformación y cuantización

A pesar de la reducción de entropía que supone la compensación de movimiento, la secuencia de errores de predicción presenta una fuerte estructura (alta correlación entre las muestras).

Usualmente se resuelve esta limitación cuantizando la señal a codificar en un determinado dominio de representación. Este es el método utilizado directamente en compresión de imágenes (donde no existe fase de compensación de movimiento).

Los algoritmos convencionales de compresión de imágenes (o de la secuencia de error) se basan en la cuantización de conjuntos de muestras (de regiones) de la imagen. De esta forma cada región cuantizada queda expresada mediante un representante discreto perteneciente a un conjunto de vectores de reconstrucción. Para aprovechar las relaciones entre los puntos vecinos, se utilizan aproximaciones vectoriales en el dominio espacio-temporal, o aproximaciones escalares, coeficiente a coeficiente, en algún dominio transformado[1]. En primera instancia, el objetivo del cambio de dominio de representación es eliminar las dependencias entre los coeficientes para justificar la cuantización escalar de la transformada [3, 8–10].

La selección de la base sobre la que expresar la señal se basa en un análisis en componentes principales de la familia de imágenes a codificar. Mediante este tipo de análisis se halla la transformación lineal que pasa del dominio inicial al dominio de ejes propios de la distribución de puntos de entrenamiento: la transformada de Karhunen-Loève (KLT) [3, 7, 9]. Además de descorrelacionar los coeficientes de la representación, la KLT presenta otras propiedades muy adecuadas para la codificación de señales[2]. Además de su caracter lineal (limitado), el problema con la KLT es que depende de la estadística del conjunto de

[1] La cuantización escalar de una señal en un dominio transformado de otro equivale a una determinada cuantización vectorial en el dominio original [9].

[2] La KLT organiza las funciones base según su relevancia en términos de error de truncamiento y minimiza dicho error para una longitud de la expansión dada, es decir, minimiza el error de representación para un número limitado de características [7, 9].

entrenamiento y hay que recalcular las funciones base en cada ocasión. Afortunadamente se ha encontrado que las funciones base de la KLT tienden a las funciones base de la transformada de coseno (DCT) si la autocorrelación de las señales en el dominio espacial tiene unas ciertas características que son razonablemente ajustadas para regiones estacionarias de las imágenes naturales [8].

Desde un punto de vista más amplio, la descorrelación de los coeficentes de la transformada no es el único objetivo de la transformación de la señal. Si se utiliza una transformada respecto de unas funciones base que tengan significado cualitativo, cada coeficiente representará una determinada característica de la escena, (p.ej. bordes, texturas o velocidades). Esta representación de la señal permite diseñar de forma más intuitiva un esquema de cuantización diferente para cada coeficiente en función de las necesidades de la aplicación, representando con más precisión ciertas características y ahorrando esfuerzo de codificación en otras [9]. Así mismo, si la representación transformada es interpretable intuitivamente, es posible disponer de varios alfabetos de codificación aplicables en función de las características de la señal en el dominio transformado [61].

En este sentido, las transformaciones *wavelet*, subbanda o multiresolución se han mostrado muy efectivas para representar texturas y patrones periódicos [21–28], fenómenos locales [62] o movimiento [29–33]. En este tipo de transformaciones, la base está formada por funciones oscilantes enventanadas, con una localización simultánea en el dominio espacial (temporal) y frecuencial. Además, uno de los argumentos esgrimidos a favor de las representaciones multi-resolución en codificación es que dan más libertad en la selección de la base, permitiendo ajustarse a los cambios locales de la estadística de la señal de modo que se cumplan las condiciones en las que las transformaciones frecuenciales tipo DCT son aproximadamente óptimas en el sentido de la KLT [63–65].

Los estándares de compresión de imágenes y vídeo más extendidos, JPEG [66, 67], MPEG [68] y H.263 [69] utilizan la DCT 2D de bloques de tamaño fijo, aplicando una codificación diferente para cada coeficiente en función de su significado cualitativo.

Una vez elegido el dominio de representación, el cuantizador viene descrito por la densidad de vectores de reconstrucción en ese dominio [70]. La particularidad de un cuantizador escalar es que dicha función densidad es separable. En el caso de la cuantización escalar el conjunto de vectores de reconstrucción es simplemente el producto cartesiano de los niveles de cuantización definidos en cada eje propio del dominio. El problema global del diseño escalar, incluye dos problemas relacionados, primero definir la forma de la distribución de los niveles de cuantización en cada eje, y segundo, determinar el número efectivo de niveles de cuantización asignados en cada eje o coeficiente de la transformada [9]. La solución clásica [71, 72] resulta de minimizar la distorsión euclidea promedio inducida por el cuantizador de cada coeficiente. El cuantizador óptimo final está determinado por las densidades de probabilidad de los coeficientes y sus varianzas [9].

Estos cuantizadores óptimos tienen dos tipos de inconvenientes. En primer lugar, igual que ocurre con la transformación óptima KLT, los cuantizadores óptimos convencionales dependen de la estadística de las imágenes a codificar, y resulta costoso actualizarlos en un entorno de estadística cambiante. Por otro lado, es obvio que en aplicaciones en las que el destinatario último de la señal decodificada es una persona, la medida de distorsión empleada en el diseño del cuantizador debe tener sentido perceptual [52]. Se ha planteado la introducción de métricas perceptuales en el diseño convencional, pero, por un lado, esto no resuelve el problema de la dependencia del cuantizador con la estadística de las imágenes a tratar, y, por otra parte, se ha comprobado que aun introduciendo métricas perceptuales, la estadística puede desviar los resultados del diseño produciendo ocasionalmente acumulaciones de los niveles de cuantización perceptualmente indeseables [12, 73].

Los estándares actuales [66, 67], pensados para un comportamiento robusto en un amplio rango de aplicaciones utilizan un cuantizador más rígido[3], introduciendo heurísticamente la sensibilidad frecuencial del SVH a nivel umbral [74]. En el caso de vídeo [10, 68, 69] se aplica un cuantizador espacial (2D) tipo JPEG para cada fotograma de la secuencia de errores.

Algunas de las mejoras que se han propuesto en la cuantización de la DCT consisten en introducir otras propiedades espaciales además de la sensibilidad frecuencial umbral [75, 76], así como introducir aspectos perceptuales temporales para secuencias [52, 77].

1.2 Modelos de procesado de contrastes en el SVH

Los elementos básicos de un modelo del procesado visual humano al nivel que estamos tratando aquí (modelos de detección y discriminación de patrones espacio-temporales simples [19, 30–32, 34, 35, 37–39, 78]), son los siguientes:

- Cambio de representación a un dominio conjunto espacio-frecuencia, dado por una batería de detectores lineales pasa-banda 2D o 3D, con campo receptivo localizado simultaneamente en el dominio de posiciones y el de frecuencias [2, 13, 17, 18, 47, 79]. La respuesta de cada uno de estos detectores representa la presencia de un determinado patrón frecuencial local de una cierta orientación en una cierta posición [20, 22, 80, 81].

- Aplicación de una función no lineal sobre las respuestas de cada uno de estos detectores. Las no-linealidades dependen básicamente de la amplitud de la respuesta de cada canal (del contraste de cada componente del estímulo de entrada) [82, 83], pero también existe una cierta influencia, de segundo orden, de las respuestas de otros canales [37, 39, 78].

- Prescripción para el cálculo de distancias en el último dominio de representación. Dados dos patrones en la representación final, la diferencia

[3]Elegido un cierto reparto relativo del numero de niveles de cuantización por coeficiente, se utilizan cuantizadores uniformes para cada dirección, y sólamente se permite una variación global (para toda dirección) del paso de cuantización mediante un factor multiplicativo.

perceptual entre ellos se obtiene mediante una sumación de Minkowski de las diferencias en cada dimensión (frecuencias y posiciones) [39, 76]. Considerando que la sumación de las diferencias sobre frecuencias (orientaciones y niveles de resolución) es previa a la sumación sobre las posiciones, la sumación (intermedia) sobre frecuencias proporciona una evaluación local de la diferencia perceptual entre los patrones considerados. La sumación espacial posterior da la contribución global de cada una de esas contribuciones locales.

Las limitaciones de la discriminación del SVH unidas al cálculo de distancia perceptual, establece la ecuación del lugar geométrico que ocupan los patrones justamente discriminables de uno dado. Este es el elipsoide de discriminación entorno a un punto del espacio de representación.

Todos los patrones cuya representación cae dentro del mismo elipsoide de discriminación son perceptualmente indistinguibles, luego el SVH efectúa una discretización de un espacio de representación continuo de forma análoga a los cuantizadores de la transformada empleados en codificación. Watson y Daugman [14, 15, 84] pusieron de manifiesto esta analogía. En particular Watson [84] propuso *de forma cualitativa* un cuantizado perceptual de las respuestas de unos filtros tipo Gabor asumiendo unas no linealidades como las de Legge [82, 85], y analizó el funcionamiento de este codificador biológico en términos de bits por muestra.

1.3 Contribuciones y estructura de la memoria

En este trabajo, partimos de la analogía del cuantizador perceptual con el objetivo de proponer expresiones explícitas para su adecuada descripcción. Mediante esta formulación explícita es posible estudiar las diferencias cualitativas entre dicho cuantizador perceptual y los cuantizadores resultantes de un análisis convencional de la señal.

Así mismo, las descripciones del proceso de simplificación de la señal en el SVH y de la geometría del dominio transformado han sido aplicadas para la optimización perceptual de ciertos algoritmos utilizados en compresión de imágenes y video. Además de las consecuencias directas de la formulación propuesta sobre la cuantización, también han sido explotadas las restricciones útiles que impone sobre la estimación de movimiento.

Las contribuciones concretas del trabajo realizado son de tres tipos:

- Nueva formulación de la eliminación de redundancia en el SVH y consecuencias sobre las medidas de diferencia entre imágenes.

- Nuevo criterio para el diseño de cuantizadores para compresión de imágenes.

- Nuevo criterio, basado en la entropía perceptual de una señal, para el control de la estimación adaptativa de flujo óptico en compresión de vídeo.

Nueva formulación de la eliminación de redundancia en el SVH y consecuencias sobre las medidas de diferencia entre imágenes

En la llamada *aproximación de alta resolución*, cuando el número de vectores de reconstrucción es suficientemente grande [9, 70, 71], el comportamiento de un cuantizador queda definido por la densidad de vectores de reconstrucción en el dominio de representación. Esta densidad, eventualmente no uniforme, representa el grado de precisión que utiliza el cuantizador para la representación de las señales situadas en las diferentes regiones del dominio.

En el capítulo 2 proponemos una función densidad en un plano de frecuencias y amplitudes para caracterizar la distribución de percepciones justamente discriminables por el SVH. Se trata de una distribución uniforme de niveles de cuantización según una métrica perceptual (no euclidea) del dominio transformado. Esta función densidad es una forma de expresar conjuntamente la solución a los dos problemas que se plantean en el diseño de un codificador escalar de la transformada[4].

Es sencillo relacionar dicha función densidad con la métrica del dominio transformado y con las no linealidades perceptuales post-transformada [86]. Hemos comprobado experimentalmente que el uso de dicha función para el cálculo de diferencias subjetivas entre imágenes obtiene buenos resultados para una variedad de tipos de distorsión (Publicación I). Para bajas amplitudes la expresión propuesta para esta función tiende a la Función de Sensibilidad al Contraste (CSF) clásica [87, 88] (despreciando los términios no lineales), con lo que, otras metricas propuestas en la literatura [36, 89, 90], basadas en el modelo lineal de la CSF, se obtienen como caso particular a partir de la formulación presentada. En el apartado 2.2 se plantea el efecto de las no linealidades sobre el reparto de niveles de cuantización en el plano frecuencias y amplitudes, pero se deja para el capítulo 3 la comparación exhaustiva de los cuantizadores perceptuales (lineales o no lineales) con los cuantizadores diseñados según criterios convencionales, basados en la estadística de la señal (Publicación III).

La formulación presentada puede plantearse en cualquier espacio de representación frecuencial[5], aunque estrictamente debería aplicarse este tipo de razonamientos en el caso de transformadas espacio-frecuencia como las utilizadas en los modelos de percepción al uso. En el apartado 2.3 (Publicación II) se trata el problema técnico de la obtención de la función de pesos sobre cada coeficiente, aproximación umbral de la métrica propuesta, en un dominio realista (de Gabor) a partir de las expresiones de caracterizaciones análogas en otros dominios.

Nuevo criterio para el diseño de cuantizadores de imágenes

El capítulo 3 recoge las aplicaciones en codificación de imágenes y video de la medida de distorsión propuesta y la caracterización de la eliminación de

[4]El problema de la asignación de información por coeficiente y el problema de la distribución 1D de los niveles de cuantización en cada eje [9].

[5]De hecho en las Publicaciones I, III-VI, se asume un dominio DCT local (en subbloques) como en las referencias [74, 76, 91].

redundancia en el SVH.

De acuerdo con la interpretación de las limitaciones en la discriminación como una cuantización del espacio de representación, en el apartado 3.1 (Publicación III), se propone un criterio alternativo para el diseño de codificadores de imágenes que tiene más sentido perceptual que el criterio convencional. Este nuevo criterio, independiente de la estadística de las señales a codificar, se basa en limitar el máximo error perceptual posible en cada coeficiente del dominio transformado, lo cual resulta equivalente a utilizar el cuantizador perceptual propuesto en el capítulo anterior. El nuevo criterio se compara con el criterio convencional de diseño basado en la minimización del error promedio, tanto desde el punto de vista de la distribución del esfuerzo de codificación en el plano de frecuencias y amplitudes como desde el punto de vista de la aplicación, comparando las imágenes reconstruidas a iguales tasas de compresión. Los resultados muestran que el criterio propuesto con métrica no lineal proporciona mejores resultados que el criterio convencional aunque emplee la misma métrica.

Se demuestra que la especificación (heurística) JPEG sobre un cuantizador basado en la CSF es óptima según el criterio presentado en el caso particular de despreciar las no linealidades en amplitud. Como es lógico, la utilización de un modelo más completo de la eliminación de redundancia por parte del SVH incluyendo las no linealidades en amplitud mejora los resultados de JPEG (Publicaciones III y IV).

Nuevo criterio de control de la estimación de movimiento para compresión de vídeo

Los métodos de compensación de movimiento utilizados por los estándares actuales de compresión de vídeo [68, 69] y por los modos básicos de las técnicas de nueva generación [92–94] se basan en la estimación del flujo óptico [95–97].

Como señalamos anteriormente, en las aplicaciones de codificación existe una relación de compromiso entre el esfuerzo dedicado a la compensación de movimiento y a la codificación de la señal de error, porque, en estas aplicaciones lo que debe de conseguirse es una reducción conjunta del volumen de flujo óptico y error codificado para un nivel de distorsión dado [12, 59, 60, 98].

En el apartado 3.2 (Publicación V) se plantea un método iterativo para alcanzar el equilibrio óptimo entre la adaptación local del flujo óptico y el tamaño de la señal de error codificada mediante un cuantizador perceptual. En este caso se encuentra que, además de resolver el problema (cuantitativo) del compromiso entre los dos procesos, el flujo óptico jerárquico adaptado mediante un criterio perceptual presenta aspectos (cualitativos) de interés general. Según el método propuesto el flujo óptico es más robusto (Publicación V) y facilita las tareas de segmentación basada en movimiento (Publicación VI).

Esto es así porque el criterio para la adaptación local del flujo pesa de manera selectiva los diferentes canales de frecuencia, con lo que, en la práctica se tiene un criterio dependiente de la escala. La dependencia tipo pasa-banda del cuantizador centra el interés en los movimientos perceptualmente significativos, dando lugar a una descripcción del movimiento más compacta.

En el apartado 3.3 (Publicación VI) las mejoras en la estimación de movimiento y las mejoras en cuantización (incluyendo aspectos temporales) se utilizan conjuntamente para plantear un esquema de compresión de video perceptualmente eficiente. Los resultados muestran que este esquema ofrece mejores resultados subjetivos que los esquemas equivalentes que no incluyen estos factores perceptuales. Los resultados sugieren que el factor que más afecta a la mejora en la calidad de la señal reconstruida es el algoritmo de cuantización.

Capítulo 2

Eliminación de redundancia en el sistema visual humano

En este trabajo, siguiendo los modelos de detección y discriminación de contrastes [19, 30–32, 34, 35, 37–39, 78], asumimos que en una primera fase, a bajo nivel, la percepción humana consiste en un conjunto de cambios de representación de la señal que transforman la imagen de entrada definida en el dominio espacial a un dominio de características frecuenciales locales y posteriormente a un dominio de respuestas a dichas características:

$$\textit{Respuesta} \;\equiv\; \mathbf{A} \xrightarrow{T} \mathbf{a} \xrightarrow{R} \mathbf{r} \tag{2.1}$$

donde, $\mathbf{A} = \{A_x\}_{x=1}^{m}$, es la representación de la señal en el dominio de posiciones (luminancia en m posiciones); cada una de las componentes de $\mathbf{a} = T(\mathbf{A})$, $\mathbf{a} = \{a_p\}_{p=1}^{n}$, representa la respuesta del filtro sensible a cada una de las n características, p (banda de frecuencias en una cierta posición espacial); y la representación final, $\mathbf{r} = R(\mathbf{a})$, con $\mathbf{r} = \{r_p\}_{p=1}^{n}$, resulta de una transformación no lineal del vector de coeficientes $\mathbf{a}$.

En este capítulo proponemos una descripción explícita de un cuantizador perceptual del dominio de características que sea consistente con las propiedades de discriminación de patrones del SVH en dicho dominio. El cuantizador que proponemos es un cuantizador escalar del conjunto de respuestas de los filtros, $\mathbf{a} = T(\mathbf{A})$, con una distribución uniforme de los vectores de codificación según una medida de distancia pereptual en ese dominio.

En primer lugar (apartado 2.1, Publicación I y [86]), proponemos una medida de distancia en el dominio transformado relacionada con las características de la transformación R. Las condiciones de validez de dicha métrica justifican el tratamiento escalar del cuantizador. La métrica resultante (basada en datos experimentales de umbrales incrementales de contraste) se utiliza para definir la distribución de vectores del cuantizador perceptual (apartado 2.2 y Publicación III). La consideración de las no-linealidades de R para amplitudes supraumbrales implica no-uniformidades en la distribución de los vectores de codificación.

Sin embargo, si se asume un modelo simplificado lineal (si se generaliza el comportamiento umbral para todo el rango de amplitudes), la distribución de los vectores es uniforme y su densidad en cada dirección queda definida por la función de pesos que da la caracterización lineal (por ejemplo la CSF en el dominio de Fourier). En el último apartado de este capítulo (apartado 2.3, Publicación II) se trata el problema de la obtención de esta caracterización lineal en distintos dominios de representación, en particular en un dominio de Gabor (con más sentido biológico).

2.1 Métrica del dominio transformado

Los modelos más recientes de discriminación de contrastes [37, 39, 78], definen la diferencia perceptual entre dos imágenes, $\mathbf{r}$ y $\mathbf{r}+\Delta\mathbf{r}$, en el dominio de respuestas:

$$D(\mathbf{r}, \mathbf{r}+\Delta\mathbf{r}) = \left(\sum_{p=1}^{n} \Delta r_p{}^{\beta} \right)^{1/\beta} \tag{2.2}$$

Esto implica una métrica euclídea del dominio de respuestas y una sumación de orden β de los incrementos en cada dimensión p. Es posible proponer una medida de distancia en el dominio transformado, con una métrica $W(\mathbf{a})$ y una sumación de orden q, que proporcione una distancia entre $\mathbf{a}$ y $\mathbf{a}+\Delta\mathbf{a}$, compatible con la distancia de las respuestas correspondientes, $R(\mathbf{a})$ y $R(\mathbf{a}) + \nabla R(\mathbf{a}) \cdot \Delta\mathbf{a}$:

$$D(\mathbf{a}, \mathbf{a}+\Delta\mathbf{a}) = \left(\sum_{p,p'} \Delta a_p{}^{q/2} W(\mathbf{a})_{pp'} \Delta a_{p'}{}^{q/2} \right)^{1/q} = \left(\sum_{p} (\nabla R(\mathbf{a}) \cdot \Delta\mathbf{a})_p{}^{\beta} \right)^{1/\beta} \tag{2.3}$$

Es evidente que la métrica en el dominio transformado debe poderse expresar en función del gradiente de la respuesta. Asumiendo el modelo de respuesta de Watson [39],

$$R(\mathbf{a})_p = \frac{a_p{}^u}{B + \sum_{p'} H_{pp'} a_{p'}{}^v} = \frac{a_p{}^u}{P_p(\mathbf{a})} \tag{2.4}$$

donde B es una constante; u y v son los exponentes excitatorios e inhibitorios de la respuesta; y la matriz $H_{pp'}$ define la interacción entre las salidas de los distintos filtros, tenemos que los elementos de la matriz gradiente, son:

$$\nabla R(\mathbf{a})_{pp'} = \frac{dR(\mathbf{a})_p}{da_{p'}} = u \cdot \frac{a_p{}^{u-1}}{P_p(\mathbf{a})} \cdot \delta_{pp'} - v \cdot \frac{a_p{}^u \cdot a_{p'}{}^{v-1}}{P_p(\mathbf{a})^2} \cdot H_{pp'} \tag{2.5}$$

Con las aproximaciones que enumeramos a continuación, obtenemos una métrica sencilla que justifica el tratamiento escalar (dimensión a dimensión) que vamos a tomar en lo que sigue:

- Asumimos que no existe *enmascaramiento* cruzado entre respuestas diferentes, a_p, es decir, $H_{pp'} = H_p \delta_{pp'}$. Esto implica (ecs. 2.4 y 2.5) que el gradiente de la respuesta es diagonal, y además cada elemento de la diagonal depende exclusivamente de la amplitud del coeficiente en esa dimensión, $\nabla R(\mathbf{a})_{pp} = \nabla R(a_p)_{pp}$.

Esta suposición no es muy restrictiva. Mientras que en [39] se asume que la interacción entre canales de diferente frecuencia (y orientación) es pequeña y que la interacción entre distintas posiciones espaciales tiene poca extensión, en otros trabajos dicha interacción es directamente despreciada [34, 76, 84, 91].

- Asumimos sumación vectorial de las distorsiones unidimensionales, es decir $\beta = 2$ y $q = 2$, la llamada *aproximación de observador ideal* [17, 39], que es utilizada en ciertos modelos [19, 38].

  Esta suposición es más restrictiva[1] pero aquí asumiremos esta aproximación para obtener una expresión más intuitiva para la métrica, y porque no modifica sustancialmente los razonamientos posteriores (que podrían hacerse análogamente utilizando los índices de sumación correspondientes).

Asumiendo la sumación vectorial, la relación entre la métrica en el dominio transformado y la métrica (euclídea) del dominio de respuestas es una simple relación entre tensores [86]:

$$W(\mathbf{a}) = \nabla R(\mathbf{a})^T \cdot \nabla R(\mathbf{a}) \tag{2.6}$$

Si la sumación fuese de otro orden, siempre podría obtenerse W en función de ∇R aplicando algúna técnica numérica a partir de la ecuación 2.3.

Asumiendo además la ausencia de enmascaramiento entre canales diferentes, la métrica resulta ser diagonal, donde cada elemento, W_{pp}, de la diagonal es exclusivamente dependiente de la amplitud de la imagen de entrada en su propia dimensión p:

$$W(\mathbf{a})_{pp} = \nabla R(a_p)_{pp}{}^2 \tag{2.7}$$

Con todo esto, es evidente que para cualquier punto, $\mathbf{a}$, la frontera de discriminación perceptual entorno al mismo[2] será un elipsoide orientado según los ejes del dominio. Las anchuras del elipsoide de discriminación en cada dirección[3], $\Delta a_p^\star(\mathbf{a})$, dependerán del elemento diagonal $W(\mathbf{a})_{pp}$ y por lo tanto, sólo dependerán de la amplitud, a_p, en esa dimensión. Utilizando un orden de sumación superior, tendríamos cuádricas de diferente convexidad (tendiendo a paralelepípedos) en lugar de elipsoides [99], pero por lo demás el comportamiento sería el mismo.

El caracter separable de las regiones de discriminación entorno a cualquier punto del dominio y la independencia de los umbrales incrementales en cada

[1] Se han propuesto valores de $\beta = 4$ para la sumación sobre frecuencias [39] y valores aun mayores para la sumación espacial [76].

[2] Es decir, el lugar geométrico de los puntos que equidistan perceptualmente de $\mathbf{a}$ una cantidad umbral constante τ.

[3] Las *mínimas diferencias perceptibles* ó JNDs (del inglés *Just Noticeable Differences*), también llamadas *umbrales incrementales* de la variable a_p.

dirección respecto del valor de las otras dimensiones del estímulo justifican la caracterización escalar (dimensión a dimensión) del comportamiento del SVH en el dominio transformado.

Con las suposiciones realizadas, dada una función base de parámetro p con una amplitud a_p, la mínima diferencia perceptible en la dirección p, $\Delta a_p^\star(a_p)$, será aquella para la que la diferencia entre a_p y $a_p + \Delta a_p$ alcance el umbral de discriminación τ:

$$D(a_p, a_p + \Delta a_p^\star(a_p))^2 = W(a_p)_{pp} \cdot \Delta a_p^\star(a_p)^2 = \tau^2 \tag{2.8}$$

con lo cual, es posible determinar empíricamente la métrica del dominio transformado (o equivalentemente, la pendiente de la respuesta R) a partir de resultados experimentales de umbrales incrementales de amplitud de las funciones base (Publicaciones III, VI y [86]):

$$W(a_p)_{pp} = \tau^2 \Delta a_p^\star(a_p)^{-2} \tag{2.9}$$

Restringiendo nuestro estudio a la discriminación de patrones periódicos localizados con una posición fija[4], es posible definir la métrica a partir de los resultados clásicos de umbrales incrementales de redes sinusoidales de Legge [76, 82, 84], o de los resultados de experimentos similares, más exhaustivos, realizados recientemente en nuestro laboratorio [83].

A partir de los resultados de estos experimentos [83], se propuso la introducción de dependencias frecuenciales en los parámetros de la expresión exponencial clásica de Legge. Sustituyendo la expresión empírica propuesta (Publicación I) en 2.9 se tiene[5]:

$$W_f(a_f) = \tau^2 \left(S_f{}^{-1} + \frac{\frac{a_f}{L}\left(k_f\left(\frac{a_f}{L}\right)^{n_f} - S_f{}^{-1}\right)}{(k_f S_f)^{-1/n_f} + \frac{a_f}{L}} \right)^{-2} \tag{2.10}$$

donde L es la luminancia promedio local, S_f es la función de pesos para cada coeficiente (que caracteriza el comportamiento lineal del SVH a nivel umbral[6]), y k_f y n_f son funciones de la frecuencia (en ciclos/grado) ajustadas para reproducir más fielmente los resultados experimentales:

$$k_f = -0.079 \log_{10} f + 0.323 \tag{2.11}$$

$$n_f = 0.840 \frac{f^{1.7}}{0.545 + f^{1.7}} \tag{2.12}$$

[4] Considerando $p = (f, x_0)$ con x_0 constante, es decir, prescindiendo de la sumación sobre posiciones espaciales.

[5] Sustituyendo el parámetro genérico p por f (frecuencia) y poniendo un solo índice f para el elemento diagonal de la métrica.

[6] En el caso de una representación de Fourier, la función de sensibilidad umbral, S_f, es la clásica CSF que puede ser computada explícitamente mediante las expresiones de Kelly [88], Nill [89] o Nygan [100, 101].

Es interesante resaltar que la expresión propuesta para la métrica contiene un *término lineal* independiente de la amplitud, que coincide con el filtro lineal que caracteriza la detección de las funciones base a nivel umbral, y un *término no lineal* dependiente de la amplitud. El término no lineal, despreciable para bajas amplitudes, crece con la amplitud, con lo que los valores de la métrica se reducen para altos contrastes. El término dependiente de la amplitud da cuenta del efecto de (auto) enmascaramiento (que implica una menor sensibilidad para altas amplitudes) y de las no linealidades de la respuesta del sistema.

Si se desprecia la corrección dependiente de la amplitud (si se asume el comportamiento umbral que ocurre cuando $a_f \to 0$), se obtiene como caso particular una métrica basada en la función S_f (la CSF) como la propuesta por Nill [89, 102] o Saghri et al. [90].

Llamaremos *métrica no lineal* a la métrica dependiente de la entrada representada por la expresión 2.10 íntegra (considerando la dependencia en amplitud). Llamaremos *métrica lineal* a la métrica independiente de la entrada, basada en la función filtro umbral, que resulta de simplificar la expresión 2.10 despreciando el término no lineal.

En resumen, en este trabajo, proponemos como medida de diferencia perceptual entre dos patrones locales $\mathbf{A}$ y $\mathbf{A} + \Delta\mathbf{A}$, la expresión:

$$D(\mathbf{A}, \mathbf{A} + \Delta\mathbf{A})^2 = T(\Delta\mathbf{A})^T \cdot W(T(\mathbf{A})) \cdot T(\Delta\mathbf{A}) = \sum_{f=1}^{n} W_f(a_f)\, \Delta a_f^2 \quad (2.13)$$

donde W viene dada por 2.10, ya sea en su versión no lineal general o en la aproximación lineal. Esta expresión supone que la transformación T (aplicación de un banco de filtros) es lineal, pero puede utilizarse en casos más generales sustituyendo simplemente $T(\Delta\mathbf{A})$ por $\nabla T(\mathbf{A}) \cdot \Delta\mathbf{A}$.

En la Publicación I se recoge la evaluación experimental de una medida no lineal de diferencia entre imágenes de la forma 2.13 basada en nuestros datos sobre umbrales incrementales de contraste [83]. Los resultados muestran un buen acuerdo de las predicciones del algoritmo con la apreciación subjetiva de distorsión expresada por los observadores para diferentes tipos de distorsión (véanse las rectas de ajuste de la distorsión subjetiva experimental en función de la medida de distancia el algoritmo, Publicación I). En los casos analizados, la medida propuesta se comporta mejor que otras medidas propuestas en la literatura [103–108], y en particular mejor que la aproximación lineal [89, 90, 102].

2.2 Cuantización escalar uniforme del dominio transformado

Un cuantizador general (vectorial) [9, 109, 110] de un dominio de n dimensiones es una transformación, Q, que asigna a cada vector del dominio, $\mathbf{a}$, un vector de reproducción, $\mathbf{b}_i = Q(\mathbf{a})$, perteneciente a un conjunto finito de tamaño N llamado *alfabeto de reproducción*, $B = \{\mathbf{b}_i; i = 1, \cdots, N\}$. El cuantizador está completamente descrito mediante el alfabeto de reproducción, B, y una partición, $\mathcal{R}$, del dominio en N regiones. Cada una de las regiones, $\mathcal{R}_i$, está formada por los puntos a los que se les asigna el i-ésimo vector de reproducción, $\mathcal{R}_i = \{\mathbf{a} \,/\, Q(\mathbf{a}) = \mathbf{b}_i\}$.

En la aproximación de *alta resolución* [70] (asumiendo que $N \to \infty$ y que $Vol(\mathcal{R}_i) \to 0$), la forma concreta de las regiones de cuantización y las posiciones de los vectores de reproducción dentro de dichas regiones no tiene tanto interés como la densidad de vectores de reproducción en el dominio, $\lambda(\mathbf{a})$. En general esa densidad será no uniforme, de forma que el cuantizador representará con más precisión unas zonas del dominio que otras.

Un cuantizador escalar de un dominio de n dimensiones es un cuantizador particular, resultante del producto cartesiano de n cuantizadores 1D diferentes en cada uno de los ejes, p, del dominio. En este caso, es evidente que las regiones de cuantización $\mathcal{R}_i$ serán paralelepípedos orientados según los ejes y que el tamaño del alfabeto será $N = \prod_{p=1}^{n} N_p$.

La descripción de un cuantizador escalar de un dominio n-dimensional tiene, en general, dos partes [9]:

- Especificación (dimensión a dimensión) de *la forma* de cada uno de los cuantizadores 1D. En la aproximación de alta resolución esta especificación consiste en dar las densidades 1D de niveles de cuantización en cada eje, $\lambda_p(a_p)$ [71].
- Asignación relativa de niveles de cuantización por dimensión (especificación de N_p) [72, 111].

Una forma conveniente de representar los dos aspectos de la descripcción de un cuantizador escalar es definir una *superficie densidad* de niveles de cuantización, $\Lambda_p(a_p)$, en el *plano* de parámetros y amplitudes, mediante

$$\Lambda_p(a_p) = N_p \cdot \lambda_p(a_p) \tag{2.14}$$

Conociendo la interpretación del plano de parámetros y amplitudes que estemos utilizando (en nuestro caso frecuencias y contrastes), las no uniformidades de dicha función representarán intuitivamente el comportamiento del cuantizador sobre las posibles señales de entrada (Publicación III). La superficie densidad caracteriza completamente el cuantizador (en la aproximación de alta resolución) ya que, $\lambda_p(a_p)$ y N_p, pueden obtenerse trivialmente a partir de $\Lambda_p(a_p)$:

$$\lambda_p(a_p) = \frac{\Lambda_p(a_p)}{\displaystyle\int \Lambda_p(a_p)\, da_p} \tag{2.15}$$

$$N_p = \int \Lambda_p(a_p)\, da_p \tag{2.16}$$

Los conceptos referidos hasta aquí son los parámetros mediante los que se describe un cuantizador. Otra cuestión diferente (que trataremos con detalle en el apartado 3.1, Publicación III) es el método (criterio de diseño) que se siga para llegar a definir estos parámetros.

Si se utiliza un cuantizador para representar las señales de un dominio contínuo, dado un determinado vector *observado*, $\mathbf{b}_i$, no es posible saber qué punto, $\mathbf{a} \in \mathcal{R}_i$, provocó dicha observación. Así mismo, si dos entradas diferentes pertenecen a la misma región de cuantización quedarán representadas por el mismo vector de reproducción, es decir, serán *indistinguibles* para el cuantizador.

El caracter discreto de la representación, $\mathbf{b}_i \in B$, implica la reducción de la su variabilidad respecto de la señal de entrada, $\mathbf{a}$, es decir, se reduce la cantidad de información que puede proporcionar dicha representación [55]. Además, la representación de los puntos $\mathbf{a}$ mediante los elementos del alfabeto, $\mathbf{b}_i$, introduce en general una distorsión irreversible.

Las propiedades genéricas de detección y discriminación de amplitudes (contrastes) en el dominio transformado por parte del SVH [78, 82, 85, 88, 112–115] pueden interpretarse mediante un modelo de cuantizador que codifica el contínuo de amplitudes posibles mediante un conjunto discreto de percepciones de amplitud justamente discriminables. De hecho, en algún trabajo reciente sobre discriminación de contrastes [116], sin ninguna relación con la analogía del cuantizador, se ha propuesto una expresión para el número de percepciones discretas de contraste para cada frecuencia, N_f.

Aunque, evidentemente, no es posible establecer una analogía completa entre las propiedades de discriminación del SVH y un cuantizador vectorial del dominio transformado[7], sí que puede resultar conveniente describir la distribución de percepciones justamente discriminables mediante los parámetros empleados en la descripción de cuantizadores (como por ejemplo la densidad, $\lambda(\mathbf{a})$) o medir la entropía de la caracterización discreta resultante para razonar cuantitativamente, aunque sea de manera relativa, sobre la cantidad de información retenida por el SVH en su representación transformada.

La idea de un cuantizador perceptual (basado en la diferente discriminación de contrastes para las diferentes frecuencias) como modelo para analizar el com-

[7] En el SVH las fronteras de discriminación se establecen para cada tarea particular de comparación, no preexisten rígidamente como en el caso de la partición $\mathcal{R}$. Por ejemplo, considerando, por una parte, un estímulo de enmascaramiento cualquiera, $\mathbf{a}_1$, en un experimento de discriminación, ese punto se convierte automáticamente en el *centro* de la región de discriminación de *radio* JND que se estudia. Si, por otra parte, tomamos el estímulo $\mathbf{a}_2 \neq \mathbf{a}_1$ perteneciente a la región de discriminación con centro en $\mathbf{a}_1$ (y por lo tanto perceptualmente indistinguible de $\mathbf{a}_1$), y realizamos un experimento de discriminación de patrones distorsionados similar, obtendremos una región de puntos perceptualmente indistinguibles de $\mathbf{a}_2$ *no completamente solapada* con la región correspondiente a $\mathbf{a}_1$, lo cual es incompatible con la definición de las regiones de la partición de un cuantizador.

portamiento del SVH a bajo nivel ya ha sido utilizada previamente [14, 15, 84]. Sin embargo, en estos casos el cuantizador perceptual no se formuló adecuadamente (en términos de $\lambda(\mathbf{a})$ para un cuantizador vectorial o en términos de N_p y $\lambda_p(a_p)$ ó $\Lambda_p(a_p)$ para un cuantizador escalar) de manera que fuese posible la comparación directa de ese cuantizador perceptual con un algún otro cuantizador genérico definido según un criterio de diseño arbitrario.

En este trabajo representamos las propiedades de discriminación del SVH mediante un cuantizador escalar *perceptualmente uniforme* del dominio frecuencial local dado por la transformación T. El objetivo es obtener expresiones explícitas para $\Lambda_p(a_p)$ ó λ_p y N_p.

Proponemos un cuantizador perceptualmente uniforme para tener una distribución de vectores de reproducción similar a la distribución de percepciones justamente discriminables (que distan unas de otras una cantidad perceptual constante –JND–). Como hemos visto en el apartado anterior, la métrica perceptual del dominio transformado no es euclídea y por lo tanto, un cuantizador perceptualmente uniforme tendrá una distribución no uniforme de vectores de reconstrucción.

Hemos visto que, bajo ciertas condiciones, la métrica es una matriz diagonal con elementos, W_p, que dependen exclusivamente de la amplitud de la entrada para ese parámetro a_p. Estas características de la métrica propician unas regiones de discriminación separables y orientadas según los ejes del dominio, lo cual justifica la aproximación escalar.

Para obtener un cuantizador escalar uniforme según la métrica, W, tendremos que exigir que la distancia perceptual entre dos niveles de cuantización cualesquiera, $b_{p\,j}$ y $b_{p\,j+1}$, en cualquier eje, p, sea constante. Como la distancia (euclídea) entre dos niveles de cuantización para una cierta amplitud, a_p, está relacionada con la densidad 1D en ese punto y el número de niveles asignados a dicho eje,

$$\Delta b_p(a_p) = \frac{1}{N_p \cdot \lambda_p(a_p)} \tag{2.17}$$

la distancia perceptual será,

$$D(b_{p\,j}, b_{p\,j+1})^2 = \frac{W_p(a_p)}{N_p^2 \cdot \lambda_p(a_p)^2} \tag{2.18}$$

La exigencia de distancia perceptual constante para toda a_p implica que $\lambda_p(a_p)$ debe ser:

$$\lambda_p(a_p) \;=\; \frac{W_p(a_p)^{1/2}}{\displaystyle\int W_p(a_p)^{1/2} da_p} \tag{2.19}$$

Sustituyendo esta densidad perceptual para cada eje, la exigencia de una distancia fija, $D(b_{p\,j}, b_{p\,j+1})^2 = k^2$, para todas las direcciones p implica un reparto de niveles por coeficiente según:

$$N_p = \frac{1}{k} \int W_p(a_p)^{1/2}\, da_p \tag{2.20}$$

Comparando las ecuaciones 2.19 y 2.20 con las ecuaciones 2.15 y 2.16 es evidente que en este caso,

$$\Lambda_p(a_p) = \frac{1}{k} W_p(a_p)^{1/2} \tag{2.21}$$

Un cuantizador perceptualmente uniforme está, pues, completamente determinado por la métrica del dominio. Utilizando una métrica particular tendremos definido un modelo concreto de cuantizador perceptual. La figura 2.1 muestra la forma de $\Lambda_p(a_p)$ en el caso de utilizar la métrica no lineal de la ecuación 2.10 y en el caso de utilizar la aproximación lineal basada en la sensibilidad umbral. En el caso no lineal se tiene una distribución no uniforme de los vectores de reproducción, mientras que en el caso lineal se tiene una distribución anisótropa, pero uniforme (cuantizadores 1D uniformes con distinto N_p). Nótese el efecto de las no-linealidades en amplitud en el reparto de niveles por coeficiente en ambos casos (figura 2.2).

En las Publicaciones I y IV se propone un modelo de cuantizador perceptual basado en una *Función de Asignación de Información* (IAF) definida (heurísticamente) como inversamente proporcional a los umbrales incrementales de contraste. Considerando que, con las aproximaciones asumidas, la métrica es inversamente proporcional al cuadrado de los umbrales incrementales (ecuación 2.9), resulta que la referida IAF es simplemente la superficie densidad, $\Lambda_p(a_p)$, que describe al cuantizador perceptualmente uniforme:

$$\Lambda_p(a_p) = \frac{1}{k}\left(\frac{\tau^2}{\Delta a_p^\star(a_p)^2}\right)^{1/2} = \frac{K}{\Delta a_p^\star(a_p)} = IAF_p(a_p) \tag{2.22}$$

2.3 Respuesta lineal del SVH bajo cambios de representación

En la mayor parte de las expresiones presentadas hemos supuesto una transformada genérica, T, sobre unas funciones, $\mathbf{G}_p = \{G_{px}\}_{x=1}^m$, con $p = 1, \cdots, n$, propias de las características p. Los razonamientos propuestos son cualitativamente válidos para cualquier tipo de representación sobre funciones oscilantes enventanadas donde, $p = (x, f)$, indica la posición de la ventana y la frecuencia espacial de la oscilación.

En la mayoría de las aplicaciones que hemos desarrollado (Publicaciones I, III, IV, V y VI), hemos aplicado propiedades del SVH determinadas mediante funciones base de la transformada de Fourier a funciones propias de otra transformada (como por ejemplo la DCT en bloques). Esta ha sido una aproximación habitual tanto en aplicaciones de ingeniería [66, 68, 74, 76], como en el desarrollo de modelos de visión humana [39, 84].

Aunque, tanto las propiedades de detección en función de la frecuencia como la variación de la sensibilidad con la amplitud son similares para todo este tipo

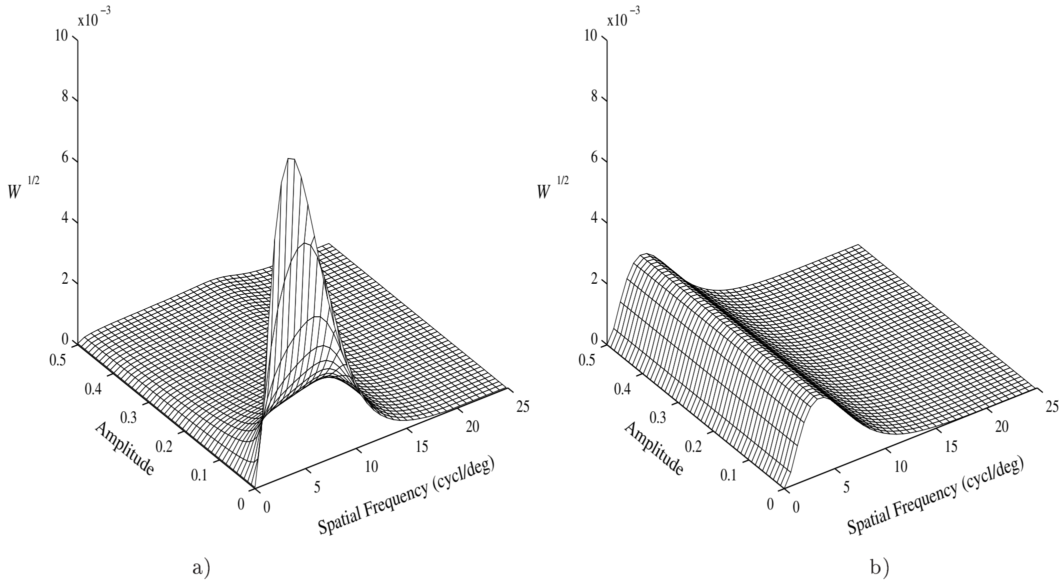


Figura 2.1: *Superficie densidad perceptual de niveles de cuantización ($\Lambda_p \propto W_p^{1/2}$) en los casos, a) métrica no lineal, dependiente de la frecuencia y la amplitud y, b) métrica lineal, exclusivamente dependiente de la frecuencia.*

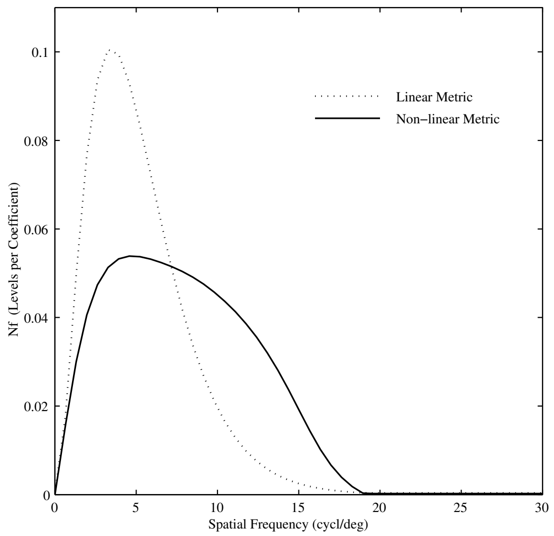


Figura 2.2: *Número relativo de niveles de cuantización por coeficiente en el caso de usar la métrica no lineal y la lineal.*

de funciones oscilantes de soporte compacto, los valores concretos de las curvas dependerán de la función base empleada. Si al proponer una métrica concreta (y el cuantizador correspondiente) hemos utilizado la expresión del filtro CSF_f y los datos de umbrales incrementales de contraste de redes sinusoidales, las expresiones obtenidas serán estrictamente válidas sólo si T es una transformada de Fourier.

En este apartado (Publicación II y [117]) tratamos el problema del cambio de la caracterización del sistema en distintos dominios de representación debidos al uso de distintas transformadas T. En particular, proponemos una técnica para calcular la respuesta lineal, S_p, a las diferentes características, p, a partir de los datos sobre dicha caracterización en otro dominio, S_f, sin necesidad de hacer experimentos sobre las funciones base del nuevo dominio $\mathbf{G}_p$, que sería lo obvio.

Aunque la respuesta del sistema visual a las diferentes caracteríaticas p extraídas mediante la transformación T no es lineal, tradicionalmente [87, 88, 112] se han utilizado funciones de pesos sobre los coeficientes de la transformada (de Fourier) como aproximación razonable del comportamiento del sistema ante estímulos cercanos al umbral.

Las caracterizaciones lineales están basadas en la determinación del umbral de detección de las funciones base de la transformada considerada, o lo que es lo mismo, en la determinación de la pendiente de R para bajas amplitudes. En este caso, se define una función de *sensibilidad* para cada coeficiente p, mediante

$$S_p = \frac{1}{\Delta a_p^\star(0)} = \nabla R(0)_p \tag{2.23}$$

En esta aproximación, se supone que la respuesta del sistema viene dada por el producto,

$$\mathbf{r} = S \cdot \mathbf{a} \tag{2.24}$$

donde, S, es una matriz diagonal con los coeficientes del filtro. Con lo que, la *imagen percibida* por el sistema es simplemente la imagen original distorsionada por la actuación de los pesos S_p:

$$\widehat{\mathbf{A}} = T^{-1}(S \cdot T(\mathbf{A})) \tag{2.25}$$

La idea para relacionar las caracterizaciones lineales en varios dominios consiste en imponer que la señal resultante de ambos procesos sea la misma independientemente del dominio donde se haya caracterizado el sistema.

Conocida la matriz S de pesos sobre los coeficientes, a_f, de la transformada T, si se quieren hallar los pesos equivalentes S' sobre los coeficientes, a_p, de otra transformada T', hay que exigir que las dos caracterizaciones den lugar al mismo resultado reconstruido para cualquier entrada, $\mathbf{A}$:

$$T'^{-1}(S' \cdot T'(\mathbf{A})) = T^{-1}(S \cdot T(\mathbf{A})) \tag{2.26}$$

En particular, utilizando como entrada las funciones, $\mathbf{G}_p$, base de la transformada T', y haciendo uso de que estas funciones son deltas en el dominio propio de esa transformación, es posible despejar cada uno de los elementos de la diagonal de S':

$$S'_p = T'\left(T^{-1}\left(S \cdot T(\mathbf{G}_p)\right)\right)_p \tag{2.27}$$

En la Publicación II esta expresión general se ha utilizado para obtener la caracterización lineal del SVH en un dominio (espacio-frecuencia 4D) de Gabor [118] a partir de nuestros datos experimentales sobre el filtro S_f en el dominio (frecuencial 2D) de Fourier [119].

Según los modelos de percepción a este nivel [19, 30–32, 34, 35, 37–39, 78], la primera etapa del análisis de la señal por el SVH es la aplicación de un banco de filtros lineales localizados en frecuencia y en el espacio, con un cierto recubrimiento del dominio de frecuencias. Esto quiere decir que tiene más sentido proponer un modelo lineal en el dominio de la transformada, T', de Gabor que en el dominio de Fourier. Mientras en el primer caso, cada valor S_p representará la atenuación introducida por el sistema en la respuesta de cada filtro p, en el dominio de Fourier, cuyas funciones base son infinítamente extensas en el espacio, y totalmente localizadas en frecuencia, los coeficientes, S_f, no tienen sentido en términos de pesos sobre ciertas señales biológicas.

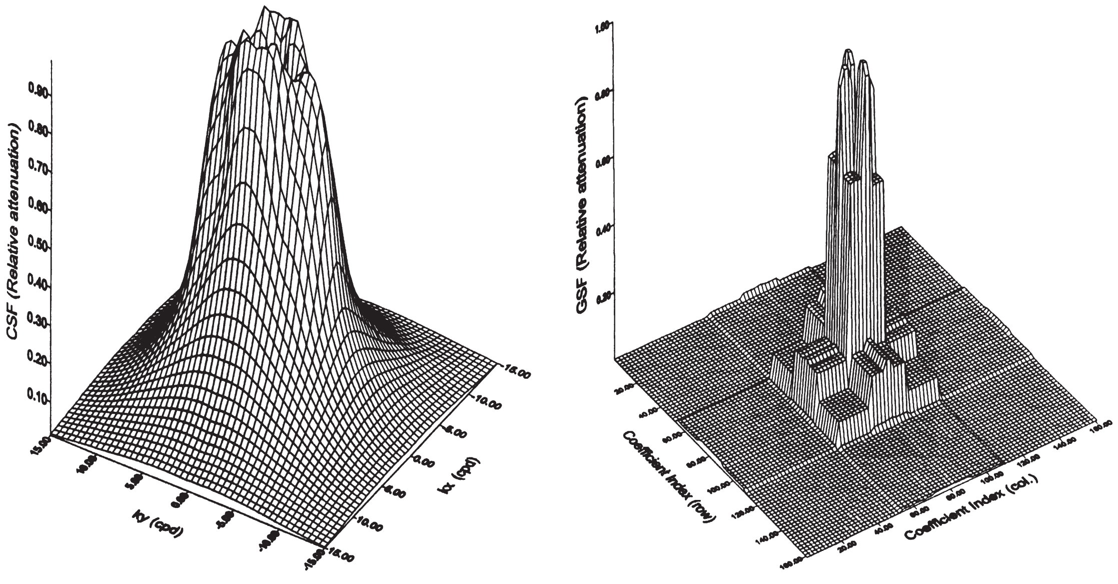
En este trabajo (Publicación II) hemos utilizado el algoritmo de Ebrahimi y Kunt [120] para el cálculo de la transformada de Gabor con una base de filtros separables y de anchura creciente con la frecuencia. En ese algoritmo los coeficientes de la transformada se obtienen mediante la minimización del error de reconstrucción [3, 121].

La figura 2.3 muestra la CSF experimental de un observador y la función de pesos equivalente en el dominio de coeficientes de la transformada de Gabor elegida. La validez de las expresiones propuestas y de las funciones filtro resultantes se demuestra mediante la similitud de las respuestas impulsionales calculadas haciendo uso de cada caracterización (Fourier y Gabor).

La caracterización del SVH en un dominio de funciones con localización simultánea en frecuencia y posición espacial [3, 79, 122] tiene interés no sólo debido a la similitud de estas funciones con la respuesta impulsional de los filtros corticales que implementan biológicamente la transformada, sino porque además, entre otras cosas, permite una caracterización sencilla de un procesado espacialmente variante.

En la Publicación II se propone un método para la caracterización o síntesis de sistemas inhomogéneos a partir de representaciones de Fourier (conocidas) de validez local combinadas en distintas posiciones del dominio de Gabor.

Aplicando la relación 2.27 utilizando diferentes funciones filtro 2D espacialmente invariantes, S, pueden obtenerse las correspondientes funciones espacialmente invariantes 4D, cuyos valores pueden combinarse en coeficientes con



a) b)

Figura 2.3: *a) Función de pesos (CSF) experimental en el dominio de Fourier. b) Función de pesos equivalente en el dominio de Gabor elegido, calculada mediante 2.27. En esta representación, la frecuencia de la función asociada a cada coeficiente crece desde el centro (como en el caso de Fourier), y el significado espacial varía de forma cartesiana en cada una de las zonas constantes de la función (en este caso el sistema es homogeneo).*

distinto significado espacial, de forma que se obtenga una sola función 4D, S', que represente el comportamiento espacialmente variante que se desee.

En la Publicación II se presenta un ejemplo en el que se sintetiza un sistema isótropo en una región espacial y anisótropo en otra. Utilizando wavelets ortogonales[8] es posible distorsionar con gran libertad las funciones base en distintas posiciones espaciales pudiéndose representar sistemas con una fuerte inhomogeneidad espacial.

La relación propuesta (ec. 2.27) constituye una solución estricta al problema que se presenta a la hora de decidir la asignación de bits por coeficiente en los algoritmos de diseño de cuantizadores para compresión de imágenes (bajo la suposición de un modelo lineal) cuando se utiliza una base de representación (DCT [66, 68, 73, 74, 123], wavelet o subbanda [120, 124, 125]) diferente de la base en la que están expresados los resultados experimentales.

[8]Las wavelets (de Gabor) utilizadas en la Publicación II no son ortogonales.

Capítulo 3

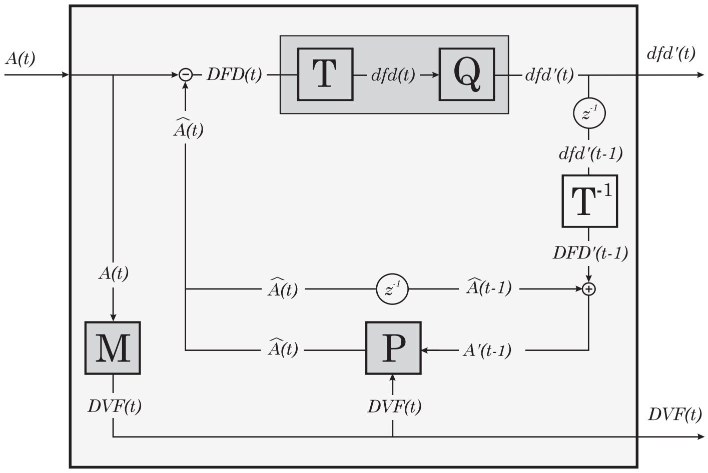
Aplicaciones en compresión de imágenes y vídeo

El objetivo general de un sistema de compresión es encontrar una expresión de la señal de tamaño mínimo para un determinado nivel de distorsión respecto de la entrada original [9, 10, 12].

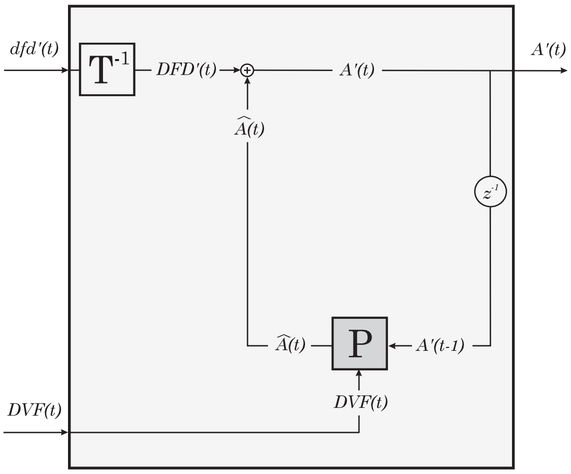
El elemento central de la mayor parte de los esquemas de compresión de imágenes y secuencias es el algoritmo de cuantización de la señal. Este proceso es el responsable de la mayor parte de la reducción del volumen de la señal codificada, así como de la introducción de errores en la señal reconstruida.

En la compresión de secuencias naturales donde la redundancia temporal de la señal es muy alta, además de una simple cuantización de la señal (3D), tiene interés utilizar algún tipo de *compensación de movimiento* antes de la cuantización. Mediante una exhaustiva descripción del movimiento de la secuencia podrían predecirse los fotogramas futuros a partir de los fotogramas previos. Sin embargo, con las técnicas actuales para la descripcción de movimiento, no es posible una reconstrucción perfecta de la señal futura a un coste (volumen de la información de movimiento) razonable. Por eso, si se utiliza una descripción compacta del movimiento es necesaria, además, una señal de corrección de los errores de predicción para consegir una reconstrucción de calidad. De esta manera, la señal original (fuertemente redundante debido a la correlación espacio-temporal entre sus muestras) se expresa mediante dos señales de menor complejidad: la información de movimiento y la señal de error residual, que (esta sí) es simplificada mediante el cuantizador correspondiente para eliminar parte de la redundancia residual que pudiera contener (figura 3.1).

En el capítulo anterior presentamos los parámetros que definen el funcionamiento de un cuantizador y propusimos una cuantización concreta como modelo del comportamiento del SVH intentando simular la distribución de percepciones justamente discriminables en el dominio transformado, pero no tratamos de las técnicas para el diseño de cuantizadores *óptimos*.



Codificador



Decodificador

Figura 3.1: *Esquema general de un codificador/decodificador de vídeo con compensación de movimiento. El codificador divide la secuencia original de fotogramas,* $\mathbf{A}(t)$*, en una secuencia de información de movimiento,* $DVF(t)$ *(flujo óptico u otra descripcción más general), y una secuencia de errores de predicción cuantizados en un dominio transformado dfd'(t). El esquema se basa en la posibilidad de estimar los fotogramas posteriores a partir de los previos y de la información de movimiento,* $\widehat{\mathbf{A}}(t) = P\left(\mathbf{A}'(t-1), DVF(t)\right)$*, de forma que el error de predicción en el instante t,* $DFD(t) = \mathbf{A}(t) - \widehat{\mathbf{A}}(t)$*, y la descripción del movimiento tienen menor complejidad que la señal original. Esta idea básica implica la presencia de un módulo de estimación de movimiento,* M*. El funcionamiento del predictor está íntimamente relacionado con el tipo de información de movimiento facilitada por* M*. El par* (T, Q) *reduce la redundancia residual de la señal de error introduciendo una distorsión,* $dfd' \neq dfd$*, que impide la reconstrucción perfecta de la señal. Los símbolos* z^{-1} *representan retrasos de la señal en una unidad de tiempo discreto.*

En el contexto de la codificación de imágenes y secuencias destinadas a observadores humanos (como en multimedia, HDTV o compresión de imágenes médicas), es importante incluir características del SVH en el diseño del sistema de codificación para para mejorar la calidad subjetiva de los resultados a una cierta tasa de compresión [52]. En este capítulo aplicamos la formulación del funcionamiento del SVH presentada en el capítulo anterior para proponer mejoras concretas a los algoritmos convencionales que se utilizan en estimación de movimiento y codificación de la transformada para compresión de imágenes y vídeo.

En el apartado 3.1 (Publicaciones III, IV y [126]), proponemos, por una parte, correcciones perceptuales a los algoritmos convencionales de diseño de cuantizadores de imágenes (basados en la estadística de las imágenes a codificar), y por otra parte, proponemos un criterio de diseño diferente (independiente de la estadística de las imágenes), más significativo perceptualmente. La calidad subjetiva de los resultados obtenidos a las mismas tasas de compresión demuestra la superioridad de la métrica no lineal frente a la lineal y del criterio de diseño propuesto frente al criterio convencional.

En el apartado 3.2 (Publicaciones V y VI) se propone utilizar la entropía resultante del cuantizado perceptual para controlar la adaptación local de una estimación de movimiento jerárquica. En los estándares más recientes de compresón de vídeo [69, 93, 94] la compensación de movimiento utiliza el flujo óptico calculado mediante correspondencia entre bloques de tamaño variable [127–132].

El uso de la entropía post-cuantizado (*entropía perceptual*) proviene, en primera instancia, de la necesidad de encontrar una relación de compromiso óptima entre el esfuerzo dedicado a la descripcción del movimiento y el dedicado a codificación de la señal de error [12, 59, 60, 98]. Sin embargo, las ventajas más interesantes del algoritmo jerárquico propuesto son la robustez del flujo resultante y la coherencia cualitativa con el movimiento de los objetos de la escena (independientes de los objetivos particulares de la codificación de vídeo).

En el apartado 3.3 (Publicación VI) se reunen las mejoras propuestas en la cuantización (incluyendo algunos aspectos temporales del SVH) y estimación de movimiento para dar un esquema de compresión de vídeo, analizándose la importancia relativa de cada contribución.

3.1 Alternativas para el diseño de cuantizadores de imágenes

El diseño convencional de cuantizadores para compresión de imágenes se basa en la minimización del promedio de una medida del error introducido por el cuantizador en un conjunto de imágenes de entrenamiento [9]. Desde este punto de vista, basado en la estadística de la clase de imágenes a codificar, la manera natural de incluir las propiedades del SVH en el diseño es la utilización de medidas de distorsión subjetiva [73, 133, 134]. En el diseño de cuantizadores escalares de la transformada usualmente se utilizan pesos dependientes del coeficiente (independientes de la entrada), S_f, basados en un modelo de percepción lineal [9, 10, 73].

Una filosofía alternativa a la aproximación basada en la estadística de la señal con correcciones perceptuales, consiste en un diseño completamente perceptual (independiente de la estadística de las imágenes a tratar) que imite la forma en como el SVH codifica la información visual [53, 135]. Desde este punto de vista, la aplicación directa de un codificador que simulase el comportamiento del SVH eliminaría directamente los datos perceptualmente irrelevantes con bajas consecuencias en la distorsión subjetiva.

Basándonos en las aportaciones del capítulo 2, en este apartado proponemos por una parte, contribuciones dentro del criterio de diseño convencional generalizando las expresiones usuales para incluir una métrica perceptual no lineal, y por otra, proponemos un criterio de diseño alternativo (la *restricción del máximo error perceptual*) con más sentido subjetivo que la *minimización del error promedio*.

Demostramos que algunos de los esquemas de cuantización recomendados (heurísticamente) en el standard JPEG son casos particulares óptimos según el criterio propuesto utilizando una métrica lineal. Se comprueba que, los cuantizadores óptimos según el criterio propuesto utilizando una métrica no lineal obtienen mejores resultados subjetivos que JPEG y que los algoritmos convencionales corregidos mediante la misma métrica percerptual. De esta forma, se comprueba la superioridad de la métrica no lineal frente a la lineal y del criterio de diseño propuesto frente al criterio convencional.

3.1.1 Minimización del error perceptual promedio

Desde el punto de vista estándard de la teoría de la tasa-distorsión[1], el problema del diseño del cuantizador para la transformada de una señal es un problema de minimización con restricciones. En función de si la restricción es el tamaño del alfabeto de reproducción [9, 10, 70–72, 111, 136, 137], o la entropía de la señal cuantizada [138, 139], se tienen distintos resultados. En este trabajo utilizamos el tamaño del alfabeto como restricción en el diseño.

El objetivo de este proceso de optimización es obtener la superficie densidad[2], que minimiza el promedio de una medida de distorsión como el error cuadrático, $\overline{D^2}$, (MSE) entre las imágenes originales y las cuantizadas.

En el diseño del cuantizador escalar de una transformada en primer lugar se resuelven los n problemas 1D, se obtienen las densidades, $\lambda_f(a_f)$, que minimizan la distorsión promedio para cada coeficiente, $\overline{D_f^2}$, y después, estas densidades

[1] Conocida en la literatura en inglés como *Rate-Distortion Theory.*

[2] O las densidades 1D, $\lambda_f(a_f)$, y el número de niveles por dimensión, N_f.

óptimas se utilizan para decidir cual debe ser el reparto de niveles por coeficiente para minimizar la distorsión total, $\overline{D^2}$.

Utilizando una medida perceptual de distorsión como la propuesta en el capítulo 2 (Publicación I y [86]), la distorsión perceptual promedio en cada dimensión es la suma de las distorsiones en cada región de cuantización:

$$\overline{D^2}_f = \sum_{j=1}^{N_f} \int_{\mathcal{R}_j} (a_f - b_{f\,j})^2\, W_f(a_f)\, p(a_f)\, da_f \tag{3.1}$$

donde $b_{f\,j}$ es el j-ésimo nivel de cuantización del eje f, $\mathcal{R}_j$ es la región de cuantización correspondiente y $p(a_f)$ es la función densidad de probabilidad (*pdf*) de la amplitud a_f.

La clave para obtener $\lambda_a(a_p)$ es la expresión de la ecuación 3.1 en función de la densidad, para obtener la llamada *integral de distorsión de Bennett* [71]. Utilizando los resultados asintóticos de Gish et al. [136] y Yamada et al. [137] para métricas no euclídeas, se tiene que la integral de de Bennett, es en este caso:

$$\overline{D^2}_f = \frac{1}{12 N_f^2} \int \frac{W_f(a_f)\, p(a_f)}{\lambda_f(a_f)^2}\, da_f \tag{3.2}$$

El cuantizador 1D que minimiza el error perceptual (no lineal) promedio puede obtenerse a partir de la desigualdad de Hölder de la forma usual [9]:

$$\lambda_{f\,MSE}(a_f) = \frac{(W_f(a_f)\, p(a_f))^{1/3}}{\displaystyle\int (W_f(a_f)\, p(a_f))^{1/3}\, da_f} \tag{3.3}$$

Este resultado difiere en el término $W_f(a_f)$ del resultado clásico proporcional a la raiz cúbica de la *pdf* [9, 70, 140, 141]. De esta manera introducimos las no linealidades perceptuales en la distribución no uniforme de los niveles de cuantización. Sustituyendo estas densidades óptimas en 3.2 se obtiene la distorsión óptima por coeficente en función de N_f,

$$\overline{D^2}_{f\,MSE} = \frac{\sigma_f^2}{12 N_f^2} \left(\int (W_f(\sigma_f a_f)\, \tilde{p}(a_f))^{1/3}\, da_f \right)^3 = \frac{\sigma_f^2}{N_f^2} \cdot H_f \tag{3.4}$$

donde $\tilde{p}(a_f)$ es la *pdf* de varianza unidad del coeficiente, y la varianza, σ_f^2, y el término H_f dependiente de la métrica, pueden considerarse como los parámetros que determinan cual es la contribución intrínseca de cada coeficiente al MSE global. La distribución óptima de niveles por coeficiente debe ser proporcional a la contribución de cada coefiente [9] (distorsión constante k^2 para cada coeficiente), por lo tanto, en este caso,

$$N_{f\,MSE} = \frac{\sigma_f}{k} \cdot H_f^{1/2} = \frac{\sigma_f}{12\,k} \left(\int (W_f(\sigma_f a_f)\, \tilde{p}(a_f))^{1/3}\, da_f \right)^{3/2} \tag{3.5}$$

A partir de estos resultados asintóticos (válidos cuando $N_f \to \infty$) la asignación de niveles por coeficiente se obtiene utilizando técnicas numéricas que aseguren un número positivo de bits por coeficente [9]. Estas técnicas están controladas por la distorsión de cada coeficiente en función del número de niveles (ec. 3.4). Por otra parte, los cuantizadores 1D en cada eje se obtienen mediante técnicas iterativas de agrupamiento[3] de N_f prototipos según el conjunto de entrenamiento [6, 109, 141].

La introducción explícita de las no linealidades del SVH a través de una métrica dependiente de la amplitud en las ecs. 3.3 y 3.5 representa un avance cualitativo en la aproximación al diseño de cuantizadores basada en la señal. Hasta el momento, se sugería una cuantización de Max-Lloyd proporcional a $p_f(a_f)^{1/3}$ [140, 141] en cada coeficiente y un reparto de niveles por coeficiente proporcional a la varianza con unos pesos frecuenciales [9, 10, 73]. En otras ocasiones, el peso de la varianza por la función de sensibilidad frecuencial, S_f, se había propuesto de forma heurística sin hacer referencia a la minimización de un error perceptual promedio [61, 89, 100, 101, 142].

A partir de la formulación propuesta, este tipo de aproximaciones se obtiene de manera natural asumiendo un modelo de percepción lineal, $W_f = S_f^2$. En ese caso, la métrica sale de las integrales de amplitud como una constante y su efecto se reduce a un peso sobre la varianza de cada coeficiente proporcional a la sensibilidad, S_f, en la expresión de N_f.

3.1.2 Restricción del error perceptual máximo

La forma natural de evaluar (empíricamente) la calidad de una imagen codificada implica una comparación *uno-a-uno* entre la versión original y la versión codificada de la imagen [108, 143]. El resultado de esta compración estará relacionado con la capacidad del observador para percibir el ruido de cuantización en presencia del patrón de enmascaramiento (la propia imagen). Esta detección o evaluación del ruido uno-a-uno está claramente relacionada con las tareas que realiza un observador en los experimentos que dan lugar a los modelos estándard de discriminación de patrones [19, 39]. En estos experimentos un observador debe evaluar la distorsión perceptual del estímulo ante desplazamientos en alguna dirección a partir de un estímulo de enmascaramiento.

Por contra, una hipotética evaluación del comportamiento global de un cuantizador sobre un conjunto de imágenes implicaría algún tipo de promediado de cada una de las comparaciones uno-a-uno. No está claro como haría tal promediado un observador humano, y además, la tarea en si misma está lejos de la comparación uno-a-uno que se establece de forma natural cuando uno mira una imagen particular.

Las técnicas convencionales de diseño de cuantizadores de la transformada aseguran la obtención de un MSE mínimo (suma de las distorsiones uno-a-uno pesadas por su probabilidad, ec. 3.1). Pero, evidentemente, la minimización del

[3] *Clustering* en la literatura en inglés.

error promedio no garantiza que en todas las comparaciones individuales vaya a obtenerse un buen resultado subjetivo [12, 144].

Aunque se utilice una métrica de distorsión subjetiva, la estadística particular del conjunto de entrenamiento puede concentrar excesivamente los vectores de reproducción en las regiones del dominio más densamente pobladas con objeto de reducir el MSE. De esta manera, ciertas regiones (con importancia perceptual) pueden quedar escasamente representadas, de forma que para los estímulos (poco probables) de esa zona se produce una fuerte distorsión perceptual. Por ejemplo, B. Macq [73] utilizó cuantizadores uniformes en lugar de los cuantizadores de Max-Lloyd porque comprobó que a pesar de minimizar el error promedio, ocasionalmente producían errores molestos en imágenes (subbloques) individuales.

Con objeto de prevenir altas distorsiones perceptuales en imágenes individuales debido a la presencia de coeficientes en regiones mal representadas, en este trabajo (Publicación III) proponemos la restricción del *error perceptual máximo* (MPE[4]) en lugar de la minimización del error perceptual promedio (MSE) como criterio de diseño del cuantizador escalar de la transformada. Esta exigencia se satisface mediante una distribución (perceptualmente) uniforme de los vectores de reproducción disponibles: si la distancia perceptual entre los niveles de cuantización en todo el rango de amplitud de cada dimensión es constante, el maximo error perceptual cometido siempre estará acotado para cualquier imagen de entrada.

Por tanto, el cuantizador óptimo según el criterio del MPE es el mismo que desarrollamos en el apartado 2.2 como modelo de la cuantización perceptual. Las expresiones equivalentes a los resultados MSE, ecs. 3.3 a 3.5, en el caso MPE son:

$$\lambda_{f\,MPE}(a_f) = \frac{W_f(a_f)^{1/2}}{\displaystyle\int W_f(a_f)^{1/2}\, da_f} \tag{3.6}$$

$$D^2_{f\,MPE} = \frac{1}{N_f^2}\left(\int W_f(a_f)^{1/2}\, da_f\right)^2 \tag{3.7}$$

$$N_{f\,MPE} = \frac{1}{k}\int W_f(a_f)^{1/2}\, da_f \tag{3.8}$$

Es interesante resaltar que, en este caso, la distribución no uniforme de los niveles de cuantización (y de los vectores de reproducción) depende exclusivamente de las características geométricas del espacio de representación, y es independiente de las propiedades estadísticas de la señal en este espacio.

La formulación MPE incluye un caso particular muy interesante en el caso de asumir la aproximación lineal de la métrica. Bajo la aproximación lineal el cuantizador MPE óptimo consiste en un conjunto de cuantizadores 1D uniformes con una asignación de niveles por coeficiente proporcional a la sensibilidad

[4] Del inglés *Maximum Perceptual Error*.

umbral, que es el esquema recomendado (heurísticamente) en los estándares JPEG [10, 66, 67, 74] y MPEG [10, 11, 68].

La formulación MPE esta cualitativamente relacionada con otras propuestas para el diseño de cuantizadores que se basan completamente en criterios perceptuales. Watson [76] propuso un algoritmo para adaptar la asignación de niveles por coeficiente a una imagen particular variando el paso de cuantización de cada coeficiente de acuerdo con la amplitud promedio de dicho coeficiente en la imagen considerada con el objetivo de obtener la misma distorsión en cada coeficiente frecuencial. También existe una versión de este algoritmo en la que, en lugar de una sumación espacial, el cuantizador se modifica para cada subbloque para mantener la distorsión total constante en todos los subbloques [91]. Por otra parte, Daly [75] propuso la aplicación de una transformación no lineal a cada coeficiente antes de la aplicación de un cuantizado uniforme (euclideo). La dependencia en frecuencia y amplitud de la sensibilidad no uniforme en la que se basaban las no linealidades aplicadas tiene una forma similar a la de nuestra superficie densidad no lineal.

3.1.3 Cuantización MSE frente a cuantización MPE

En este apartado (Publicación III) se presentan los resultados de los cuantizadores obtenidos según los diferentes criterios de diseño (MSE y MPE) y las diferentes métricas de distorsión consideradas (euclídea, métrica perceptual lineal –con dependencias frecuenciales– y métrica perceptual no lineal –con dependencias en frecuencia y amplitud–).

La figura 3.2 muestra las superficies densidad para las distintos diseños analizados. En ellas se puede ver, separadamente, el efecto de la estadística de las imágenes (que concentra los niveles en las regiones de bajo contraste), el efecto de la sensibilidad frecuencial umbral (que concentra el esfuerzo de codificación en la región de frecuencias intermedias), y el efecto de las no linealidades en amplitud (que modifica la asignación de niveles por coeficiente).

En los casos MPE, independientes de la estadística de las imágenes, la distribución está exclusivamente determinada por la métrica perceptual elegida. En el caso lineal (tipo JPEG y MPEG) la banda frecuencial de interés es muy estrecha. La consideración de las no linealidades ensancha la banda de paso y concentra el interés del cuantizador en la zona de bajos contrastes, aunque de forma menos acusada que en el caso MSE.

Es interesante comparar la forma de la superficie densidad en el caso MPE no lineal, –característica de la percepción humana–, con la superficie densidad de un cuantizador óptimo según un criterio exlusívamente estadístico como el MSE euclídeo. A pesar de la semejanza cualitativa (en ambos casos, mayor relevancia de las bajas frecuencias y de las bajas amplitudes), parece que el SVH concede mayor importancia relativa a los detalles de alta frecuencia y alta amplitud ($f > 10(cpd)$ y contraste > 0.1, –ver Publicación III–) que un cuantizador MSE.

La similitud cualitativa entre las superficies densidad explica el hecho de que los cuantizadores puramente estadísticos alcancen unos resultados razona-

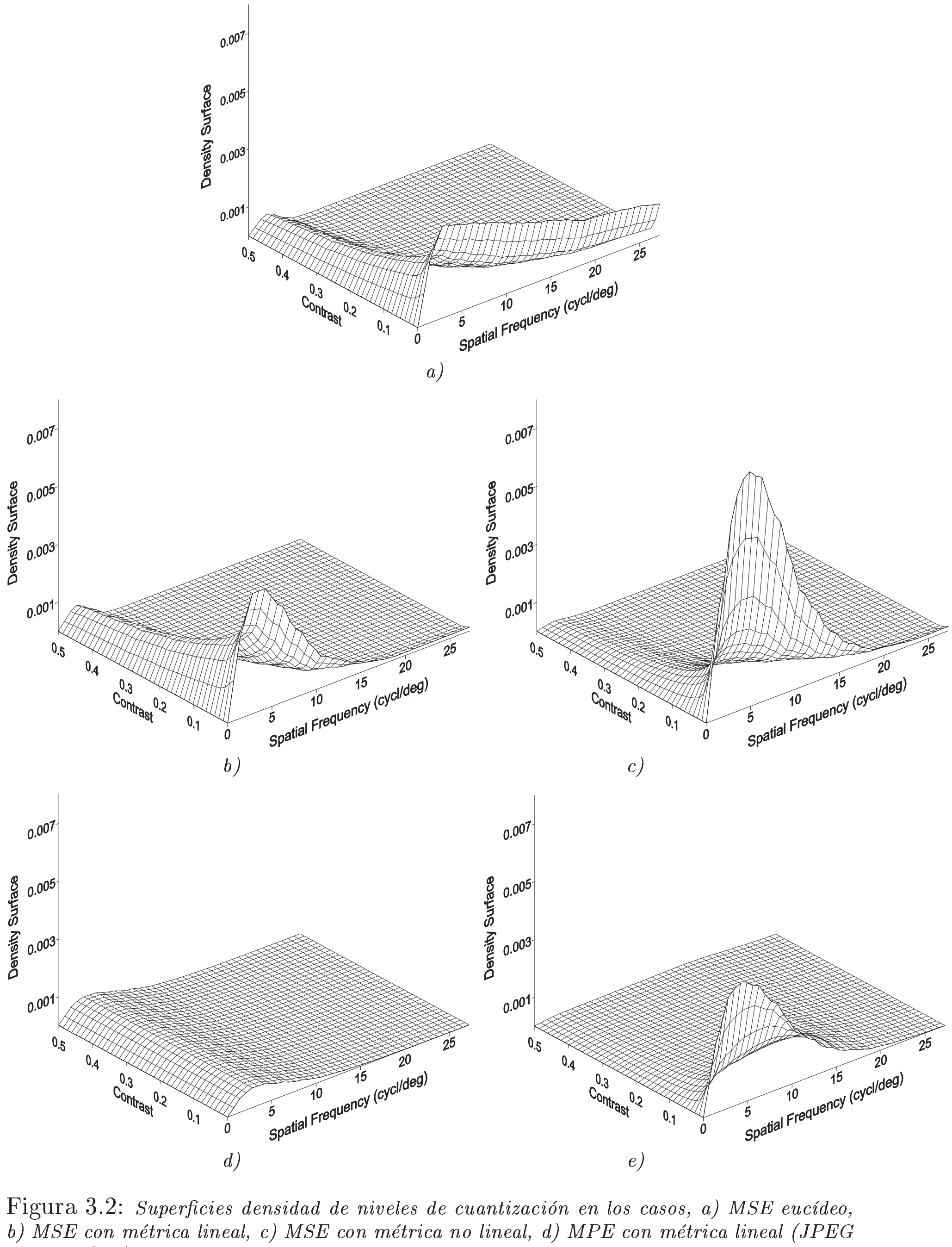


Figura 3.2: *Superficies densidad de niveles de cuantización en los casos, a) MSE eucídeo, b) MSE con métrica lineal, c) MSE con métrica no lineal, d) MPE con métrica lineal (JPEG ó MPEG), e) MPE con métrica no lineal.*

blemente buenos. Además, parece confirmar a grandes rasgos las ideas de Field y Kersten [16, 145, 146], sobre la adaptación de la respuesta del SVH a la estadística de las imágenes con las que trata.

Sin embargo, las diferencias (no despreciables) entre las superficies revela que el diseño de un compresor de imágenes según un criterio exclusivamente estadístico (al menos con una norma de orden 2) no va a obtener en general un resultado perceptualmente óptimo. Por otra parte, se pone de manifiesto que la sensibilidad del SVH en el plano de frecuencias y contrastes puede estar fuertemente influenciada por la estadística de las imágenes naturales, pero en ningún caso se trata de una relación sencilla.

En las figuras 5 y 6 de la Publicación III se muestran varios ejemplos del funcionamiento de los distintos esquemas de cuantización a la misma tasa de compresión. Los resultados de los esquemas MSE, incluso cuando utiliza una métrica dependiente de la frecuencia y la amplitud, son muy similares a los resultados del MPE lineal (tipo JPEG ó MPEG). Por contra, el esquema MPE no lineal obtiene claramente los mejores resultados subjetivos, especialmente en los detalles de alta frecuencia. Esto es así debido a la fuerte influencia de las funciones densidad de probabilidad de los coeficientes, que desequilibran las distribuciones de niveles de cuantización hacia los bajos contrastes especialmente en frecuencias medias y altas.

Mientras que las diferencias entre los cuantizadores JPEG y MPEG y el cuantizador MPE no lineal se justifican por el empleo de un modelo perceptual que incluye más aspectos del SVH (Publicaciones III y IV), las diferencias entre los cuantizadores MPE y MSE no lineales están exclusivamente basadas en el uso de un criterio de diseño distinto. Estos resultados sugieren que el criterio MSE no puede aprovechar las ventajas de métricas perceptuales más sofisticadas. Por otro lado, los resultados confirman que restringir el error perceptual máximo en cada imagen individual tiene más sentido perceptual que minimizar el error perceptual promedio sobre un conjunto de entrenamiento.

3.2 Criterio alternativo para el refinamiento local del flujo óptico

La descripcción del movimiento de la escena en codificación de vídeo tiene por objeto predecir los fotogramas futuros a partir de la secuencia previa. Las técnicas actuales de estimación de movimiento pueden englobarse en dos grandes grupos: flujo óptico [58, 95, 96], que no requiere de un modelo a priori de la escena, y técnicas basadas en la correspondencia entre características de los objetos de la escena [49, 97, 147], lo cual implica un cierto grado de modelización de la misma.

Tanto en los estándares más extendidos MPEG-1, H.261 [10, 11, 68], como en los más recientes H.263 [69] y MPEG-4 [94], se utiliza una descripción del movimiento basada en el flujo óptico calculado mediante correspondencia entre bloques en fotogramas sucesivos (BMA, del inglés *Block Matching Algorithm*) [58, 148].

Aunque las próximas generaciones de algoritmos de codificación de vídeo harán uso de técnicas de descripción del movimiento de más alto nivel [48, 50, 149–151], no se descarta el uso del flujo óptico tanto para obtener información inicial que permita la elaboración de modelos de escena en las primeras fases de análisis de la secuencia [92, 93, 152, 153], como para usarlo como modo de predicción simple y robusto en caso de fallos por desajustes del modelo complejo.

La segmentación *quadtree* proporcionada por los algoritmos BMA de resolución adaptativa que usan bloques de tamaño variable (p.ej. en H.263) representa una solución intermedia entre un cálculo ciego de flujo óptico a resolución uniforme sin ninguna relación con el contenido de la escena, y los algoritmos basados en modelos de los objetos móviles presentes en la secuencia. Este tipo de algoritmos adaptativos permite aumentar localmente el esfuerzo de estimación de movimiento centrándose en las regiones con cambios más significativos.

De esta forma, se puede explotar la relación de compromiso que existe entre el esfuerzo dedicado a la estimación de movimiento (DVF, del inglés *Displacement Vector Field*) y el dedicado a la codificación de los errores de predicción (DFD, del inglés *Displaced Frame Difference*).

En los algoritmos BMA basados en estructuras localmente adaptables como las *multigrid* o *quadtree*, la estimación de movimiento se inicia a una baja resolución (bloques de tamaño grande). En cada resolución se busca el mejor desplazamiento para cada bloque a través del cálculo de una medida de similitud (usualmente simple correlación) para un conjunto restringido de desplazamientos. La resolución de la estimación de movimiento es localmente incrementada (un bloque de la estructura se divide) si se satisface un determinado criterio de división. El proceso de estimación de movimiento finaliza cuando no es posible dividir ningún bloque del nivel inferior de la estructura.

El criterio de división es la parte más importante del algoritmo porque el grado de refinamiento de la estimación tiene efectos sobre los volúmenes relativos del DVF y del DFD [12, 59, 154], y puede dar lugar a estimaciones inestables si la resolución se incrementa innecesariamente [128]. Se han propuesto varios criterios de división para BMAs jerárquicos:

- Una medida de la magnitud del error de predicción [127–132], la energía del DFD, el error cuadrático o el error absoluto promedio.
- Una medida de la complejidad del error de predicción. En este caso, se han propuesto medidas de entropía del DFD en el dominio espacial [59, 154], o de la entropía del DFD codificado [12, 60, 98].

Los criterios basados en la magnitud de la diferencia se han propuesto sin una relación específica con los procesos de cuantización, y por lo tanto, no explotan la relación de compromiso existente entre DVF y DFD en codificación de vídeo. Se ha resaltado que los criterios basados en entropía son muy adecuados en aplicaciones de codificación de vídeo porque consiguen minimizar el volumen conjunto de DVF y DFD [59, 154]. Sin embargo, las medidas de entropía en el dominio espacial no dan buena cuenta del comportamiento de los cuantizadores pasa-banda con base perceptual que se utilizan tras la estimación de movimiento. El comportamiento pasa-banda particular del cuantizador perceptual puede introducir efectos interesantes en el criterio de división.

Se han propuesto aproximaciones rigurosas basadas en la teoría de la tasa-distorsión para obtener la asignación de bits óptima entre DVF y DFD [12, 60, 98], sin embargo, a pesar de la consideración implícita de la codificación de la transformada, los efectos cualitativos de una cuantización no uniforme de la transformada del DFD no fue analizada en estos casos.

En este apartado (Publicaciones V y VI) proponemos un criterio de división basado en la entropía perceptual de la señal de error, con la idea de refinar la estimación de movimiento sólo si el esfuerzo adicional implica una reducción de entropía perceptualmente significativa.

3.2.1 Criterio de división basado en la entropía perceptual

Un criterio de refinamiento del flujo que conceda igual importancia a todos los detalles de la señal de error es un *criterio plano*, como son los basados en la magnitud del error [128–132] o en la entropía espacial de orden cero [59, 154].

Como el DFD va a ser codificado por un cuantizador perceptual (selectivo, *no plano*), no todos los detalles mejor predichos mediante una mejor estimación de movimiento van a ser relevantes para el cuantizador posterior. Desde el punto de vista de ahorro de esfuerzo en la estimación de movimiento, no merece la pena un refinamiento de la estimación si luego el cuantizador introduce una distorsión equivalente a la mejora que se ha conseguido.

En el contexto de codificación de vídeo destinada a un observador humano, la resolución de la estimación de movimiento sólo debe incrementarse en las zonas donde el error debido a una mala predicción sea *perceptualmente significativo*. Para identificar cuáles son las áreas de interés para el cuantizador, proponemos la utilización de la *entropía perceptual* del error. De acuerdo con [14, 15, 84], llamamos entropía perceptual, H_p, de una señal, a la entropía de la representación codificada por el SVH. En nuestro caso, H_p será la entropía de la representación discreta dada por el cuantizador perceptual, Q_p, descrito en el capítulo 2, y que hemos propuesto en 3.1 para codificar imágenes:

$$H_p(\mathbf{A}) = H\left(Q_p[T(\mathbf{A})] \right) \tag{3.9}$$

Dada una cierta descripción del movimiento (mediante una cierta información $H(DVF)$), un incremento en la resolución, y por lo tanto un incremento de volumen $\Delta H(DVF)$, será perceptualmente relevante sólo si esta información adicional implica una mayor reducción de la entropía perceptual de los errores de predicción:

$$\Delta H(DVF) < -\Delta H_p(DFD) \tag{3.10}$$

Esta definición de lo que es información de movimiento (perceptualmente) relevante implica un criterio de división basado en la entropía perceptual del error:

Un bloque de la estructura quadtree debe ser dividido si,

$$H(DVF_{div}) + H_p(DFD_{div}) < H(DVF_{no\ div}) + H_p(DFD_{no\ div}) \tag{3.11}$$

donde $H(DVF)$ es la entropía del campo de vectores de movimiento y $H_p(DFD)$ es la entropía perceptual del error (con o sin división del bloque).

Esta restricción perceptual para la estimación de movimiento proviene de la aplicación particular de codificación de vídeo adaptada a las exigencias del destinatario final de la señal. Sin embargo, los beneficios de la inclusión de las propiedades de la respuesta del SVH a las diferentes características resultantes de una transformada frecuencial van más allá de la optimización del volumen conjunto del DFD y el DVF, debido a que el caracter pasa-banda de la medida de entropía perceptual implica una estrategia de división dependiente de la escala que puede ser útil para discriminar entre movimientos significativos y falsas alarmas (figuras 3.3 y 3.4).

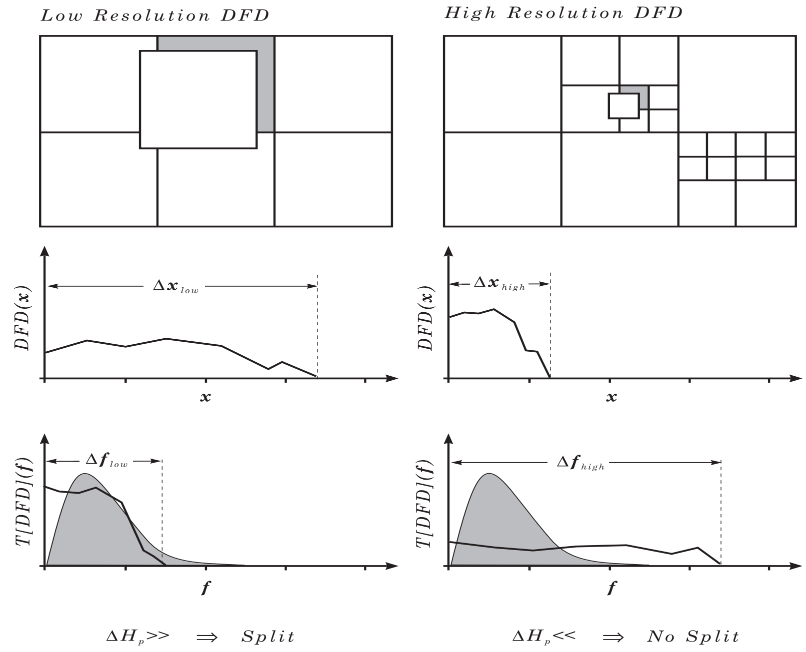


Figura 3.3: *Criterio de división dependiente de la escala debido a la realimentación perceptual. Para un nivel de energía dado, la extensión espacial y la anchura de banda del DFD están relacionados por la relación de incertidumbre, $\Delta x \cdot \Delta f = k$. Por lo tanto, la anchura de banda dependerá de la resolución dando lugar a un comportamiento diferente a diferentes resoluciones debido al carácter pasa-banda del criterio de división.*

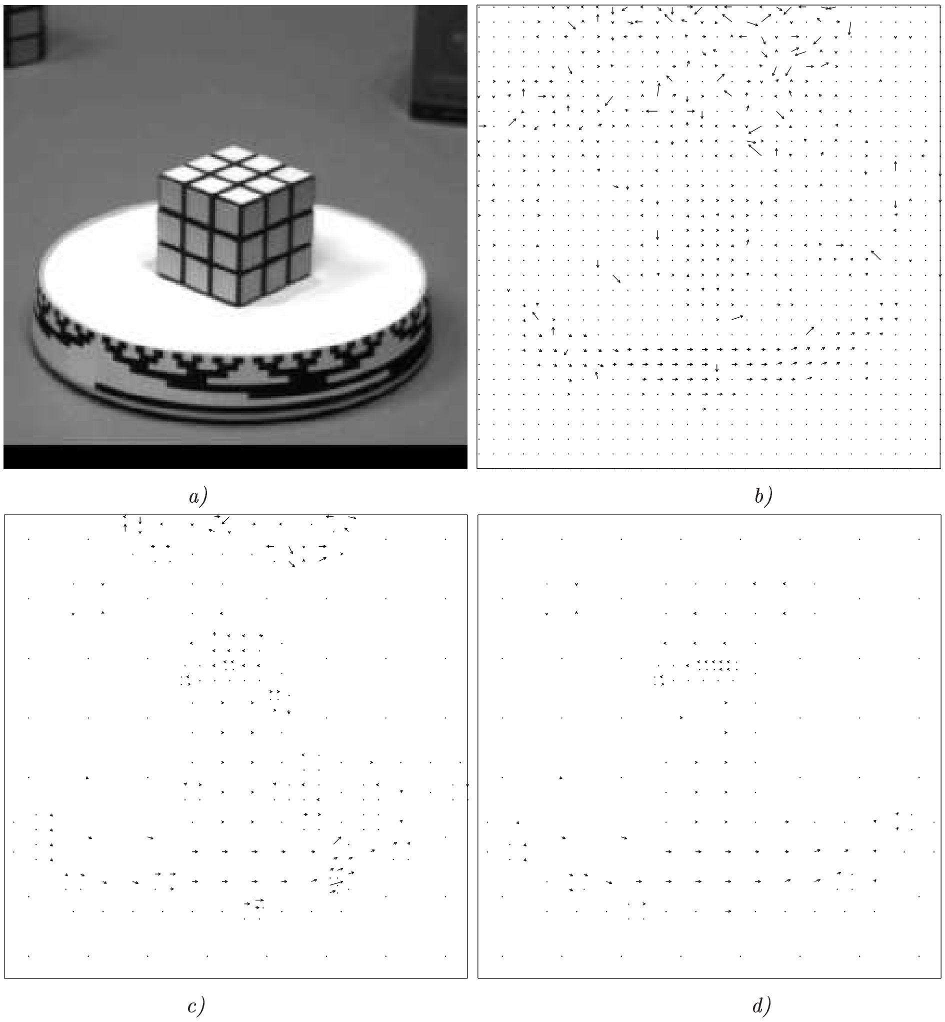


Figura 3.4: *Efectos del criterio de división en el flujo óptico estimado. a) Fotograma 6 de la secuencia RUBIK. b) Flujo óptico (fotogramas 6-7) obtenido mediante BMA de resolución uniforme. c) Flujo óptico obtenido mediante BMA adaptativo con criterio plano. d) Flujo óptico obtenido mediante BMA adaptativo con criterio perceptual.*

3.2.2 Efectos de la realimentación perceptual en la estimación de movimiento

El objetivo principal (cualitativo) de una estimación de movimiento localmente adaptativa es obtener una descripción del movimiento con mayor potencia predictiva y que, por lo tanto, elimine una mayor cantidad de redundancia temporal. El objetivo particular (cuantitativo) del control de la adaptación mediante una medida de entropía consiste en obtener una asignación de bits óptima entre DVF y DFD, con la idea de que un mejor equilibrio en esta asignación redunde en más bits para codificar el error y conseguir una mejor calidad en la señal reconstruida.

Por lo tanto, hay dos posibles aproximaciones para el análisis de los resultados de movimiento, una cuantitativa, en términos de tasa de compresión [12, 60, 98]), y otra cualitativa, en términos de la utilidad y las propiedades del flujo óptico obtenido.

Comparando el algoritmo propuesto basado en la entropía perceptual con el algoritmo *plano* basado en la medida de entropía en el dominio espacial se obtienen los siguientes resultados:

- Desde el punto de vista cuantitativo, el algoritmo propuesto (que utiliza la entropía de la señal codificada) es óptimo en el sentido de minimizar la entropía real de la señal comprimida. Este resultado es obvio por la definición del algoritmo (ec. 3.11). Cualquier algoritmo que utilice la entropía del cuantizador posterior (independientemente de su naturaleza) [12, 60, 98] hubiese obtenido un resultado similar frente al algoritmo espacial de Dufaux et al. [59, 154].

  Sin embargo, se comprueba que la reducción del volumen del flujo (del orden de un 2% respecto del algoritmo de Dufaux) no es suficiente para que el cuantizador codifique con más precisión el DFD y se obtenga una mejor reconstrucción, luego, las ventajas cuantitativas de un algoritmo adaptativo óptimo frente a un algoritmo adaptativo sub-óptimo no parecen ser relevantes.

- El interés del algoritmo propuesto no reside en el hecho de que alcanza un reparto óptimo de bits entre DVF y DFD, sino, en que obtiene un flujo óptico con mejores características cualitativas: más robusto, más preciso en la medida de velocidades y con una mejor descripción de los movimientos significativos de la secuencia.

  Estos resultados no tienen que ver con el uso de la entropía real de la señal codificada, sino con el hecho de que esta entropía es la entropía perceptual, que tiene una determinada selectividad frecuencial.

El comportamiento del algoritmo propuesto puede simplificar las tareas de interpretación de la escena a más alto nivel basadas en medidas precisas de velocidad y en segmentación de objetos basada en movimiento. Esta representación mejorada del movimiento puede ser útil como etapa inicial en un esquema de codificación de vídeo basado en modelos.

3.3 Propuestas para H.263 y MPEG-4: efecto de las mejoras en movimiento y cuantización

En este apartado (Publicación VI) proponemos un esquema de compresión de vídeo con compensación de movimiento y codificación de la DCT de la señal de error[5] completamente basado en criterios perceptuales. La idea del esquema propuesto es eliminar toda información que no sea subjetivamente relevante haciendo uso del modelo presentado en el capítulo 2.

Para ello se han utilizado cuantizadores óptimos según el criterio MPE no lineal (para eliminar adecuadamente la redundancia subjetiva del DFD) y estimación de flujo óptico jerárquico controlada mediante un criterio de entropía perceptual (que centra la atención sólo en los movimientos perceptualmente significativos).

La secuencia de errores DFD es una señal espacio-temporal y por lo tanto, en principio debe conseguirse un mejor aprovechamiento de la información disponible (una mejor eliminación de la redundancia subjetiva) si se tienen en cuenta las propiedades temporales del SVH.

Además del cuantizador MPE no lineal de la DCT 2D aplicado fotograma a fotograma, hemos estudiado un esquema 3D obtenido extendiendo nuestros datos experimentales sobre umbrales incrementales de contraste de redes espaciales [83] a redes espacio-temporales haciendo uso de la similitud de las no linealidades del SVH en ambos casos [114, 115].

Según esto, asumiendo una expresión para la métrica perceptual del dominio de frecuencias espacio-temporales, para una posición (x, t) fija, análoga a la ec. 2.10, considerando que $f = (f_x, f_t)$, y utilizando como función de sensibilidad umbral, S_f, la CSF 3D de Kelly [112, 113] podemos usar las expresiones 3.6 a 3.8 para definir un cuantizador de la DCT 3D óptimo según el criterio MPE.

Como resulta dificil la implementación de un cuantizador 3D en el bucle de predicción de la figura 3.1, en la variante 3D estudiada, aproximamos el efecto temporal de este cuantizador 3D mediante un filtro temporal con una respuesta frecuencial dada por el número de niveles por frecuencia temporal (integrando sobre todas las frecuencias espaciales).

En la Publicación VI se compara el efecto de las alternativas propuestas en la calidad de la reconstrucción. En primer lugar se analiza cada mejora (cuantización y estimación de movimiento) de forma aislada (utilizando un algoritmo estándar en el otro proceso) frente a la cuantización basada en la CSF de MPEG y estimaciones de movimiento uniformes y adaptativas con criterio *plano*. Así mismo se han comparado los siguientes esquemas globales (nuestras propuestas y dos estándares):

---

[5]Los algoritmos descritos separadamente en los apartados 3.1 y 3.2 corresponden a los módulos que integran la estructura de H.263 [69] y de ciertas propuestas de MPEG-4 [93, 150, 151, 155], con lo que pueden incorporarse en estos esquemas de manera sencilla.

- MPEG-1: BMA uniforme y cuantización MPE lineal [68].
- H.263: BMA adaptativo con criterio de división *plano* y cuantización MPE lineal [11, 59, 69].
- Los esquemas perceptuales propuestos: BMA adaptativo con criterio de división basado en la entropía perceptual y cuantización MPE no lineal (con y sin filtro temporal –caso 3D y 2D–).

Los resultados muestran que el factor que más afecta a la calidad de la reconstrucción es el cuantizador (respecto de otro algoritmo que utilice ya algún BMA jerárquico). La introducción de las no linealidades en amplitud en la cuantización mejoran notablemente los resultados (como podía preverse a partir de los resultados de codificación de imágenes estáticas). La ampliación de la banda de paso del cuantizador debida a la introducción de los aspectos supraumbrales hace que más detalles relevantes del DFD sean conservados impidiendo la rápida degradación de la secuencia que tiene lugar en el caso de utilizar el cuantizador uniforme. El cuantizador 2D propuesto reduce el aspecto granuloso de los bordes en movimiento, y el eventual filtrado temporal reduce la visibilidad de ruido impulsional en fotogramas y bloques aislados, uniformizando el movimiento de la escena con el coste de un ligero desenfoque de los objetos en movimiento.

Capítulo 4

Conclusiones

En esta tesis se ha planteado una métrica para el cálculo de distancias perceptuales entre patrones locales representados en un dominio frecuencial. Se ha relacionado la métrica perceptual en ese dominio con las no linealidades de la respuesta del SVH a las distintas funciones base de la transformada considerada y se ha comprobado experimentalmente la efectividad de la métrica propuesta.

De forma consistente con las condiciones de validez de la métrica propuesta, se han planteado expresiones explícitas para la descripción del proceso de eliminación de redundancia en el SVH mediante un cuantizador escalar de la transformada sobre una base con significado frecuencial.

Dicho cuantizador ha sido formulado mediante una superficie densidad de niveles de cuantización en el plano de frecuencias y amplitudes, de tal modo que su comportamiento puede ser comparado con el de otros cuantizadores definidos según criterios cualesquiera.

Se ha comprobado que la métrica propuesta se reduce a una caracterización lineal del SVH en el caso umbral (el cuadrado de la CSF considerando una transformada de Fourier), y se ha propuesto un método para hallar esta caracterización lineal en cualquier otra base de representación. En particular, se ha obtenido esta caracterización lineal en una representación wavelet de Gabor comprobándose la utilidad de las expresiones propuestas para establecer filtros (o métricas) espacialmente variantes. De este modo pueden representarse de forma sencilla sistemas con inhomogeneidades espaciales.

En los compresores de imágenes más extendidos tan sólo se consideraba la incorporación de pesos frecuenciales para tener en cuenta los aspectos perceptuales. En este trabajo se ha introducido el efecto de las no linealidades de la respuesta del SVH en el diseño de cuantizadores de imágenes de dos formas: primero, utilizando la métrica no lineal propuesta en la aproximación convencional que se basa en la minimización de una medida del error promedio (MSE), y segundo, proponiendo un criterio de diseño alternativo con más sentido perceptual, –la restricción del error perceptual máximo (MPE)–. Se demuestra que el criterio MPE da lugar a los cuantizadores de tipo JPEG (o MPEG) de manera natural asumiendo una métrica lineal. Por otro lado, se obtiene que la utilización de una métrica no lineal en el criterio MPE mejora sustancialmente los resultados. El criterio de diseño convencional (MSE) no consigue explotar las ventajas adicionales de una métrica más sofisticada debido al efecto distorsionador de la estadística del conjunto de entrenamiento.

El modelo de la eliminación de redundancia subjetiva en el SVH se ha aplicado para diseñar un sistema de codificación de vídeo con la idea de no conservar más información de la que puede aceptar el cuantizador perceptual.

Esta idea ha sido utilizada de dos formas en un esquema con compensación de movimiento como el H.263: primero, se ha utilizado un cuantizador MPE no lineal para reducir la redundancia subjetiva del error de predicción, y segundo, se ha propuesto un criterio de adaptación local para la estimación del flujo óptico basado en la reducción de la entropía perceptual del error de predicción. Los beneficios más relevantes del algoritmo de estimación de movimiento propuesto van más allá del objetivo (particular) ligado a la codificación –optimizar el esfuerzo entre estimación de movimiento y codificación del error–, y tienen que ver con las propiedades cualitativas del flujo final (robustez, precisión y coherencia) ligadas a la forma pasa-banda de la medida de entropía utilizada.

Apéndice A

Publicaciones

- J.Malo, A.M.Pons & J.M. Artigas. Subjective image fidelity metric based on bit allocation of the HVS in the DCT domain. *Image & Vision Computing*. Vol. 15, N. 7, pp. 535–548 (1997).
- J.Malo, A.M.Pons, A.Felipe & J.M.Artigas. Characterization of the HVS threshold performance by a weighting function in the Gabor domain. *Journal of Modern Optics*. Vol. 44, N. 1, pp. 127–148 (1997).
- J.Malo, F.Ferri, J.Albert, J.Soret & J.M.Artigas. The role of the perceptual contrast non-linearities in image transform quantization. *Image and Vision Computing*. (Aceptado Abril 1999).
- J.Malo, A.M.Pons & J.M.Artigas. Bit allocation algorithm for codebook design in vector quantization fully based on HVS nonlinearities for suprathreshold contrasts. *Electronics Letters*. Vol. 31, N. 15, pp. 1222–1224 (1995).
- J.Malo, F.Ferri, J.Albert & J.M.Artigas. Splitting criterion for hierarchical motion estimation based on perceptual coding. *Electronics Letters*. Vol. 34, N. 6, pp. 541–543 (1998).
- J.Malo, F.Ferri, J.Soret & J.M. Artigas. Exploiting perceptual feed-back in multigrid motion estimation for video coding using an improved DCT quantization scheme. *IEEE Transactions on Image Processing*. (Enviado Abril 1999).

A.1 Subjective image fidelity metric based on bit allocation of the HVS in the DCT domain

Image & Vision Computing. Vol. 15, N. 7, pp. 535–548 (1997)

ELSEVIER

Image and Vision Computing 15 (1997) 535–548

image and vision COMPUTING

Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain

J. Malo*, A.M. Pons, J.M. Artigas

Departament d'Òptica, Facultat de Física, Universitat de València, C/Dr. Moliner 50, 461 00 Burjassot, València, Spain

Received 15 July 1996; accepted 4 December 1996

Abstract

Until now, subjective image distortion measures have partially used diverse empirical facts concerning human perception: non-linear perception of luminance, masking of the impairments by a highly textured surround, linear filtering by the threshold contrast frequency response of the visual system, and non-linear post-filtering amplitude corrections in the frequency domain. In this work, we develop a frequency and contrast dependent metric in the DCT domain using a fully non-linear and suprathreshold contrast perception model: the *Information Allocation Function* (IAF) of the visual system. It is derived from experimental data about frequency and contrast incremental thresholds and it is consistent with the reported noise adaptation of the visual system frequency response. Exhaustive psychophysical comparison with the results of other subjective metrics confirms that our model deals with a wider range of distortions more accurately than previously reported metrics. The developed metric can, therefore, be incorporated in the design of compression algorithms as a closer approximation of human assessment of image quality. © 1997 Elsevier Science B.V.

Keywords: Subjective image fidelity; Non-linear perception model; DCT domain

1. Introduction

The importance of measuring image differences is central to many image and video processing algorithms.

The most straightforward application is image encoding design and optimisation. The final codebook depends on the selected metric controlling the iterative process which reduces the distance between the original and the encoded image [1,2]. It is known that in the applications judged by a human observer, Euclidean metrics such as mean square error lead to unsatisfactory results [3,4]. In those cases, the encoding procedure has to be designed to minimise the *subjective distortions* of the reconstructed signal. Characteristics of the human viewer (in fact a numerical model of its requirements and limitations) must be included to obtain a perceptually weighted metric that reproduces opinion of the observers [5]. Up to now, most successful image and video codec standards (JPEG [6], H.261 and MPEG-X [7,8]) can be qualified as perceptual oriented coders as they include certain human visual system (HVS) characteristics to match the encoding process to the observer requirements. All such techniques carry out a transform of the signal to a frequency domain (in particular DCT), not only due to the energy compactation and decorrelation properties of these transforms, but also due to the fact that HVS sensitivity is highly uneven in these domains [5,9]. An optimum frequency dependent bit allocation can be done in such a way that many transform coefficients can be safely discarded. These standards [6–8] and other compression schemes [10–12] have used an ad hoc knowledge of HVS threshold frequency response, but the amplitude stepsize non-linearities, when taken into account [10–12], are still based on statistical LBG design of the codebook using Euclidean metrics. If a reliable subjective metric were employed, the standard LBG algorithm for codebook design could be used without other ad hoc previous corrections.

The applications of subjective distortion metrics are not restricted to the design of a perceptually matched quantizer. Wherever image differences are computed and perceptual criteria have to be satisfied, this kind of measures can be applied. Two less exploited examples of this are motion estimation and adaptive noise cancellation. In region matching methods for motion estimation, the estimated displacement at a particular point is the vector that minimises *some difference measure* between a neighbourhood of the point in that frame and a neighbourhood of the displaced point in the previous frame [13]. In adaptive image enhancement through linear or non-linear filtering, the filter coefficients

* Corresponding author. E-mail: jesus.malo@uv.es.

PII S0262-8856(96)00004-2

locally change to minimise *some difference measure* between the output of the filter and the target reference [14,15].

All these techniques could be perceptually matched if a reliable subjective distortion measure were used to guide them.

The basic idea of any perceptually weighted metric is that subjective differences between two images cannot be directly derived from the *given* images, but from their *perceived version*. According to this, the way we model the information reduction process that the visual system applies to the input signals, is the key to obtain a good subjective fidelity metric.

Until now, the reported subjective distortion measurements have been based on selective weighting of image differences in the spatial or/and the frequency domain [3,4,16–21]. This selectivity has been founded on a perception model implementing at least one of these processes:

1. Application of a point-wise cube-root non-linearity on intensity to take into account the non-linear response of the photoreceptors to luminance [17,21,22].
2. Masking effects, weighting the intensity differences according to the values of the gradient, the local activity, the local contrast or the local luminance [3,4,21], to take into account the major sensitivity of HVS to deviations in sharp edges, and the masking of noise in highly textured areas.
3. Linear band-pass filtering by the visual system's threshold frequency response [16–22]. The perception models based on a linear filtering scheme assume that frequency sensitivity of the HVS is inversely proportional to the contrast detection threshold of sinusoidal patterns. The visual system frequency response, named *Contrast Sensitivity Function* (CSF) in the psychophysical literature [23–25], is empirically defined as the inverse of the contrast detection threshold of sinusoidal gratings of different frequencies. These linear and threshold models do not include suprathreshold non-linear performance of HVS [25,26].
4. Application of non-linear functions (logarithmic or square root functions) to the amplitude spectrum after the linear stage [18–20,22]. The response to the amplitude of the input, the sensitivity and the resolution step size of any detector are closely related. The Weber law states that the HVS has a logarithmic varying resolution step size for luminances (zero spatial frequencies). This implies high sensitivity for low luminances and lower sensitivity for higher luminances. This uneven amplitude sensitivity is also qualitatively true for the non-zero frequencies [25], but the actual resolution step size for amplitudes is not logarithmic [26]. The aim of the post-filtering non-linear functions applied to the spectrum is to emphasise the differences between the original and the distorted spectra in the perceptually significant areas (favouring low amplitudes). However, these logarithmic or square-root corrections are just heuristic extensions of the Weber law to the non-zero frequencies.

In this work we develop a frequency and contrast dependent metric in the DCT domain using a fully non-linear and suprathreshold perception model: the *Information Allocation Function* (IAF) of the visual system [26,27]. This approach qualitatively improves the previous perception models used for subjective distortion measurement because, due to the nature of its experimental foundations, it is not a stage-after-stage sequential model made of disconnected characteristics of the HVS, but it includes the effects from photoreceptors to post-transform suprathreshold non-linearities. The bit allocation model has been derived from experimental data about frequency and contrast incremental thresholds of HVS, and it is consistent with the reported noise adaptation of the CSF [23]. The IAF has been recently employed to guide the design of a perceptually adapted vector quantizer giving better subjective performance that JPEG standard at the same bit rate [27]. In this work, experimental results are presented that show how the distortions predicted by the IAF metric are linearly related to the opinion of the observers. We also compare the performance of our metric with other distortion measures partially based on processes 1 to 4. The experimental comparison confirms that our model deals with a wider range of distortions more accurately than these previously reported metrics.

2. Modelling bit allocation properties of the human visual system: The IAF

The existence of tolerances to changes in contrast of gratings or DCT base functions implies that the visual system maps the continuous contrast range into a finite set of discrete perceptions. Such a discrete encoding implies that perception can be considered as a redundancy removal process analogous to vector quantization. These considerations gave rise to a novel model of human visual performance focused in its bit allocation properties [26,27]. The Information Allocation Function (IAF) gives the amount of information assigned by the system to encode each area of the frequency-contrast domain. A inversely proportional dependence with the size of the tolerances of HVS to changes in that domain was proposed for this function [26,27]:

$$\mathrm{IAF}(f,C)=\frac{\mathrm{d}^2 I}{\mathrm{d}f\,\mathrm{d}C}=\frac{K}{\Delta f(f,c)\,\Delta C(f,C)}, \tag{1}$$

where $\Delta f(f,C)$ and $\Delta C(f,C)$ are the experimental frequency and contrast incremental thresholds of gratings in each point of the frequency-contrast domain. The experimental procedure to determine these incremental thresholds consists of measuring the minimum amount of frequency or amplitude deviation needed to make just discriminable two successively presented sinusoidal patterns of a given frequency and contrast [26]. The results obtained by such a technique

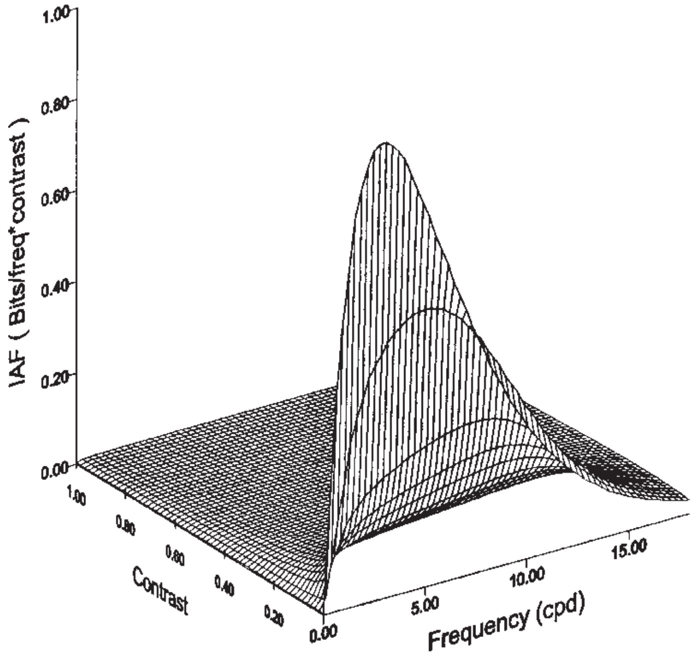


Fig. 1. Relative bit allocation of the visual system in the frequency-contrast DCT domain.

do not include spatial domain masking effects, however, it must be stressed that they do include all the possible non-linearities in the spatial and the frequency domains because they are not obtained from some intermediate stage but from the whole response of the HVS. The explicit expression for the IAF is:

$$\mathrm{IAF}(f,C) = \frac{K}{C_T(f) + C\left(\left(\frac{C_T(f)}{k(f)}\right)^{1/n(f)} + C\right)^{-1}\left(k(f)C^{n(f)} - C_T(f)\right)}, \qquad (2)$$

where $C_T(f)$ is the absolute contrast detection threshold for a grating of frequency f (the inverse of the CSF(f)) and $k(f)$ and $n(f)$ are the empirical variation of the parameters of $\Delta C(f,C)$ [26]:

$$k(f) = -0.079389 \log_{10}(f) + 0.322725, \qquad (3a)$$

$$n(f) = 0.84\frac{f^{1.7}}{0.54534 + f^{1.7}}. \qquad (3b)$$

This bit allocation model obtained from the experimental data implies higher resolution of the visual system for middle frequencies and the low contrast range (see Fig. 1), in agreement with classical perception models [24,25]. Nevertheless, quantitative comparison reveals that the IAF has a wider effective band-pass than the threshold CSF filter [24] (see Fig. 2), and that suprathreshold non-linearities are no longer a simple logarithmic or square-root transduction for each non-zero frequency. Note that considering the suprathreshold characteristics of HVS (contrast dependent term in the denominator of eqn (2)) substantially modifies the relative importance of each frequency coefficient. In other words, if one neglects suprathreshold characteristics of HVS, one gets the threshold CSF, and substantial disagreement with the opinion of the observers will be obtained. The shape of the IAF is consistent with the reported noise level dependence of the CSF [23].

3. Measuring subjective differences with the IAF

The subjective distortion measure presented in this paper is based on the fact that the amount of information assigned by the system to encode each area of the frequency-contrast DCT domain can be a good measure of the relative importance given by the system to the differences between the encoded signals in that area. The basic idea is: as more information is allocated by the HVS in one area, more visual importance is given to impairments in that area. An IAF-like frequency-contrast sensitivity surface has been successfully employed in the design of perceptually adapted lossy compression algorithms [23,27]. This fact supports the idea that the IAF could be an accurate weight to measure the contributions of differences in the DCT domain to the global distance between images.

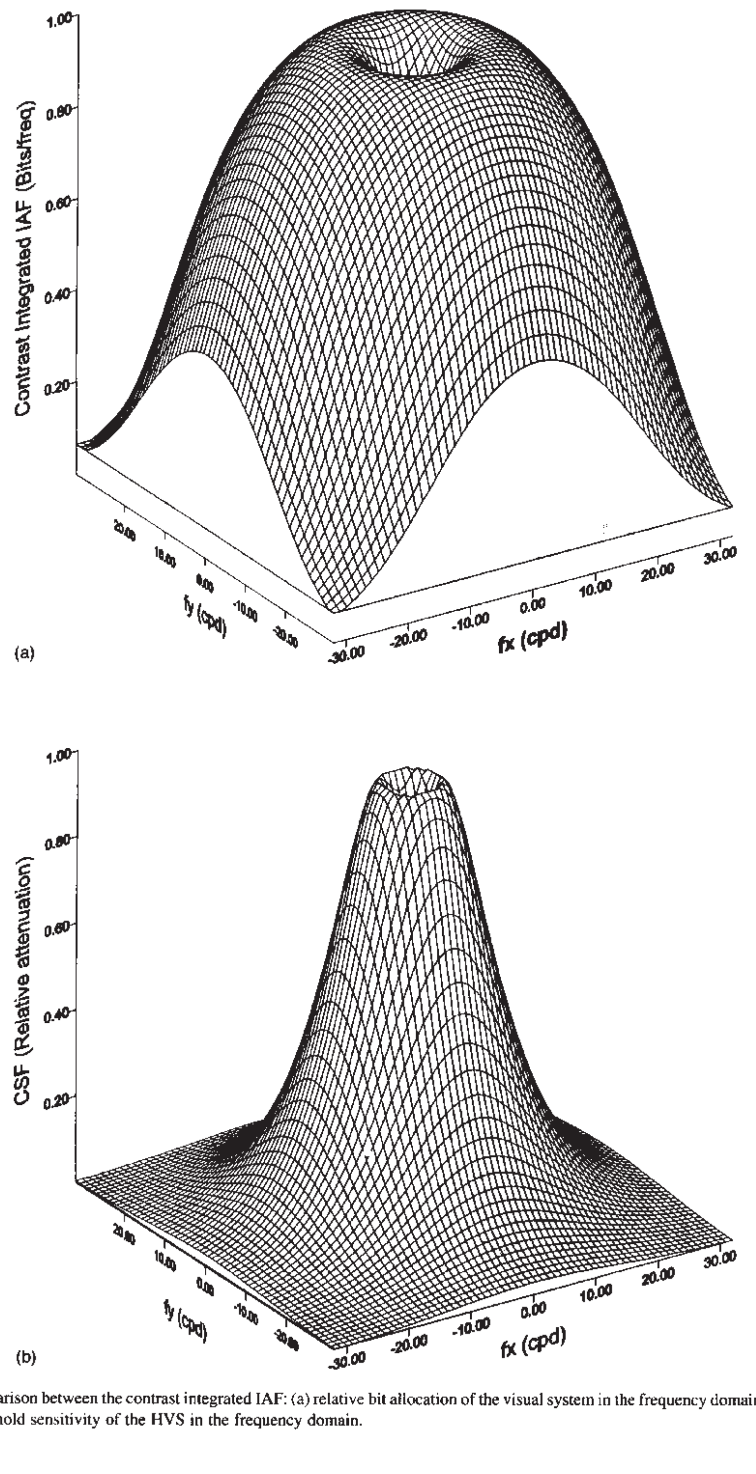


Fig. 2. A comparison between the contrast integrated IAF: (a) relative bit allocation of the visual system in the frequency domain, and (b) the CSF: the threshold sensitivity of the HVS in the frequency domain.

According to this, the measurement of the subjective difference between an original image i_1 and its distorted version i_2 can be done weighting the differences between the perceived images in the DCT domain with the IAF in that domain:

$$d(i_1, i_2)^2 = \int \mathrm{IAF}(f, C)(\mathrm{QDCT}[i_1] - \mathrm{QDCT}[i_2])^2 \, df, \quad (4)$$

where QDCT stands for the quantized DCT, f is the spatial frequency of each coefficient and C is the contrast for amplitude $(\mathrm{QDCT}[i_1] - \mathrm{QDCT}[i_2](f))/2$. Notice that in this model, perception is considered as a variable-precision quantization process given by the IAF, according to the method exposed in [13]. In this way, the integral (4) can be interpreted as the sum of the differences between the perceived images assuming the IAF model of redundancy removal.

4. Other error measurements

In this section we briefly outline some of the distortion measures reported in the literature [3,4,16–21] which are representative of the metrics that use the visual perception characteristics described in the introduction. We will use these algorithms to emphasise the quantitative improvement obtained from the IAF metric.

(1) Standard objective energy based metric: The mean square error (MSE):

$$d(i_1, i_2)^2 = \int (i_1(x) - i_2(x))^2 \, dx. \quad (5)$$

In spite of its well known disagreement with subjective assessment, standard MSE has been included here in order to analyse the nature of the distortions of the test images, and to see how the different metrics improve the straightforward result given by the MSE.

(2) Modified MSE including photoreceptors cube-root non-linear response [9]:

$$d(i_1, i_2)^2 = \int \left(i_1(x)^{1/3} - i_2(x)^{1/3}\right)^2 dx. \quad (6)$$

(3) Masking based metric I. This metric assumes that perception can be characterised by a simple blurring process (local averaging) with two additional features: there is a facilitation of error detection in sharp edge zones, and there is masking of errors in highly textured areas [3,4]. According to this, the considered metric computes the differences between the locally averaged images, weighting each difference by a factor directly proportional to the gradient of the image and inversely proportional to the local variance:

$$d(i_1, i_2)^2 = \int \left((M[i_1(x)] - M[i_2(x)]) \frac{|\nabla i_1(x)|}{1 + 2\,\mathrm{Var}[i_1(x)]} \right)^2 dx, \quad (7a)$$

where $M[i(x)]$ is the mean luminance in a neighbourhood, N, of point x:

$$M[i(x)] = \int_N i(x) \, dx \quad (7b)$$

and $\mathrm{Var}[i(x)]$ is the local variance evaluated in the neighbourhood of x:

$$\mathrm{Var}[i(x)] = \int_N (i(x) - M[i(x)])^2 \, dx. \quad (7c)$$

(4) Masking based metric II. This error measure takes into account the reduction of sensitivity in high luminance areas weighting the image differences by the inverse of the local luminance [4]:

$$d(i_1, i_2)^2 = \int \left(\frac{(i_1(x) - i_2(x))}{M[i_1(x)]} \right)^2 dx. \quad (8)$$

(5) Linear filtering based metric [16,18]. Linear filtering based models of perception assume that the perceived image is a band pass filtered version of the original image. The characteristic HVS filter is the threshold frequency response, the CSF. Assuming this, the perceptual difference should be computed from the filtered images, weighting the error spectra by the frequency response of the system:

$$d(i_1, i_2)^2 = \int (\mathrm{FT}[i_1(x)](f) - \mathrm{FT}[i_2(x)](f))^2 \mathrm{CSF}(f)^2 \, df. \quad (9)$$

(6) Non-linear photoreceptors response plus linear threshold filtering [17,21]:

$$d(i_1, i_2)^2 = \int \left(\mathrm{FT}\left[i_1(x)^{1/3}\right](f) - \mathrm{FT}\left[i_2(x)^{1/3}\right](f)\right)^2 \mathrm{CSF}(f)^2 \, df. \quad (10)$$

(7) Non-uniform weighting of spectrum differences: Subjective Quality Factor (SQF):

$$d(i_1, i_2) = \int_{f_1}^{f_2} \frac{|\mathrm{FT}[i_1(x)](f) - \mathrm{FT}[i_2(x)](f)|}{f} \, df. \quad (11)$$

(8) Non-uniform weighting of spectrum differences plus linear threshold filtering and nonlinear contrast response: Square Root Integral (SQRI) [18,19]

$$d(i_1, i_2) = \int_{f_1}^{f_2} \frac{((\mathrm{FT}[i_1(x)](f) - \mathrm{FT}[i_2(x)](f))\mathrm{CSF}(f))^{1/2}}{f} \, df. \quad (12)$$

5. Experiments

To test the reliability of the proposed metric and compare its performance with the results of the other described metrics under a variety of distortions, two different standard psychophysical experiments [21,28] have

(a) (b) (c) (d) (e) (f)

(g) (h)

(i) (j)

(k) (l)

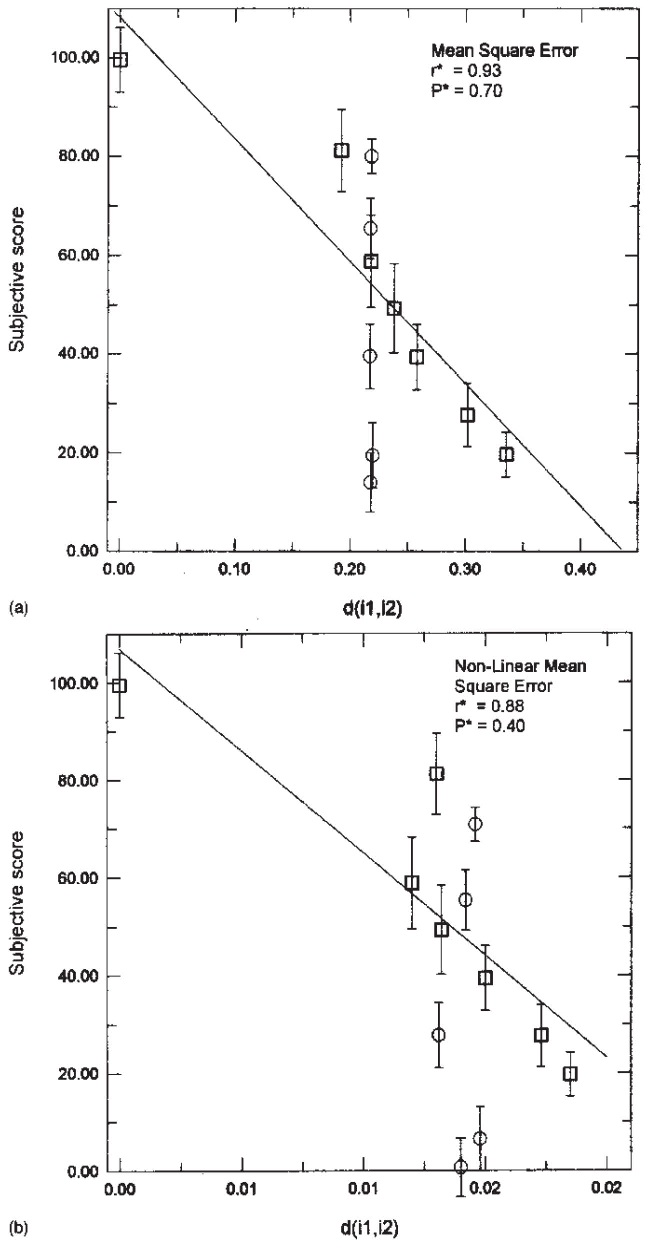
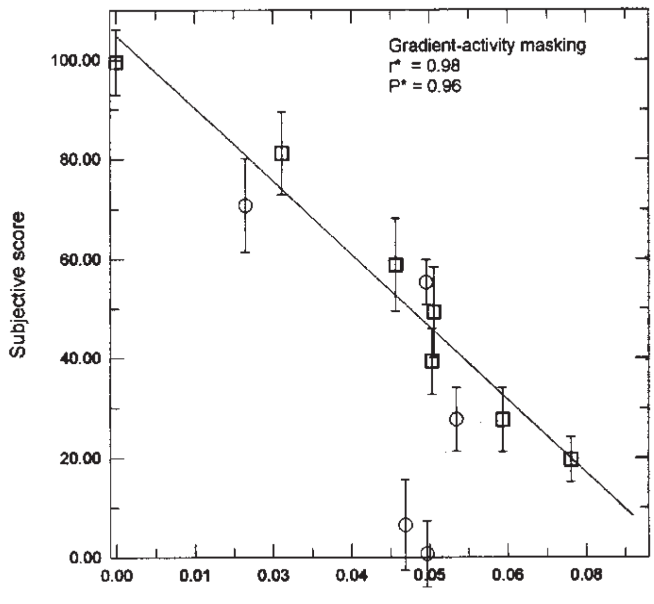
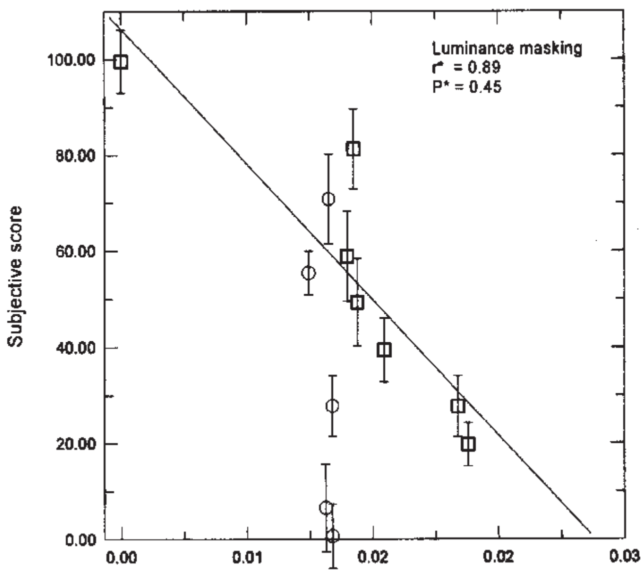


Fig. 4. Observers score vs. distortion computed by the different algorithms in experiment I (circles) and experiment II (squares). The correlation coefficient (r) and the probability of the χ^2 test (P) of the fit are given inside each figure. When they are marked with an asterisk (r^*,P^*), it means that the straight line only fits the data of experiment II. In those cases, the metric can not hold the coloured noise data (see the corresponding r and P-values for the experiment I and both experiments in Table 1).

Fig. 3. Examples of the test images used in the experiments. Experiment I, images (a–f): Original image corrupted by coloured noise of the bands [0,1] cl/deg (a); [1,2] cl/deg (b); [2,4] cl/deg (c); [4,8] cl/deg (d); [8,16] cl/deg (e); and [16,32] cl/deg (f). The signal to noise energy ratio is 100:5. Experiment II, images (g–l): JPEG-compressed images at different bit rates: 0.80 bits/pix (g); 0.38 bits/pix (h); 0.29 bits/pix (i); 0.18 bits/pix (g); 0.13 bits/pix (k); 0.08 bits/pix (l).



(c)

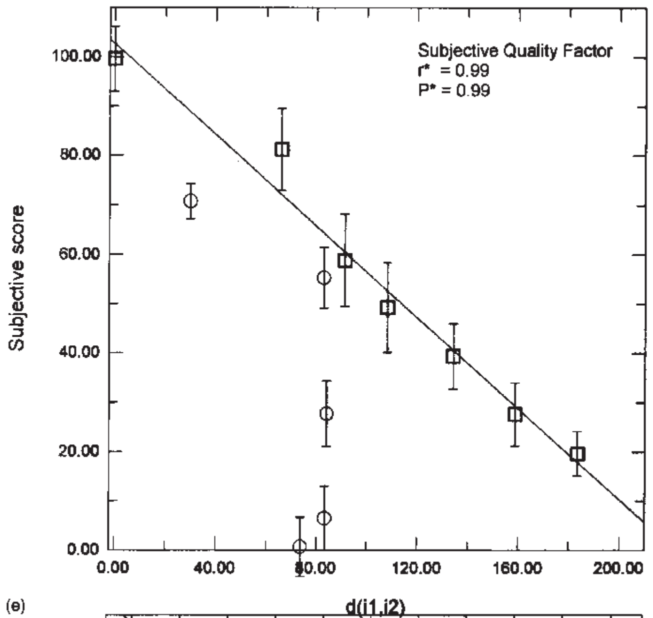


(d)

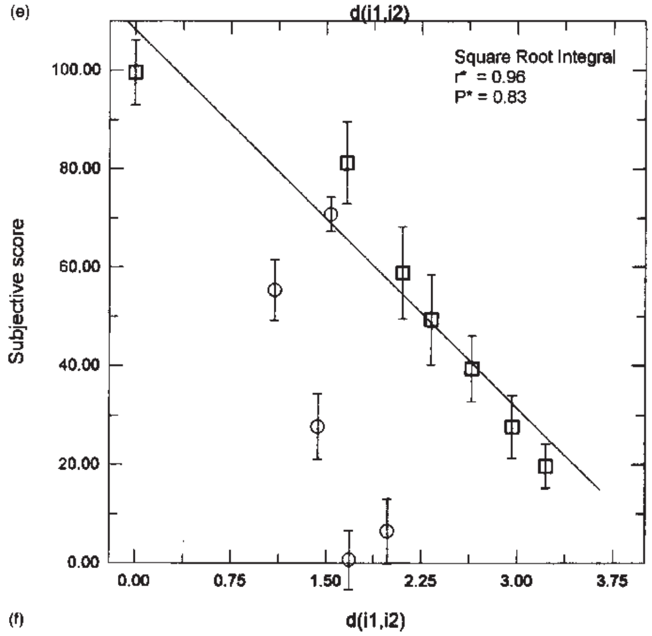
Fig. 4. Continued.

been carried out:

- Experiment I: The task of the observers was ordering a set of degraded images with constant energy noise but different spectrum, (coloured noise), so that an equal MSE would be obtained. The coloured noise was obtained filtering a white-noise random signal with a square filter of the proper band-pass. In experiment I the tests were distorted versions of a natural image



(e)

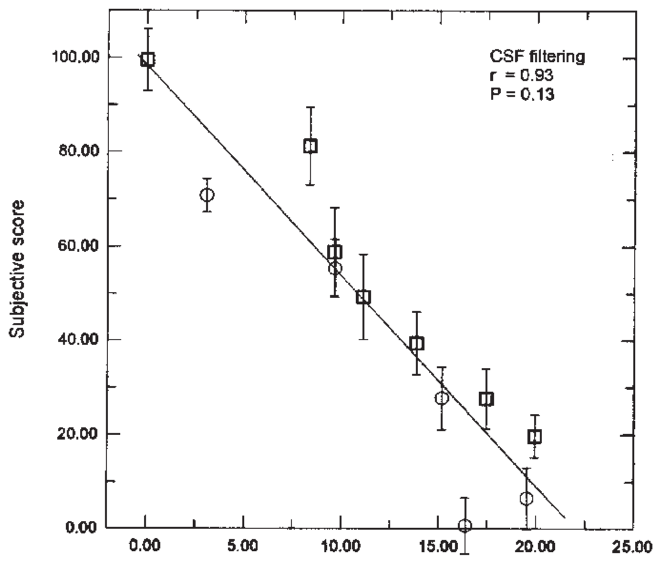


(f)

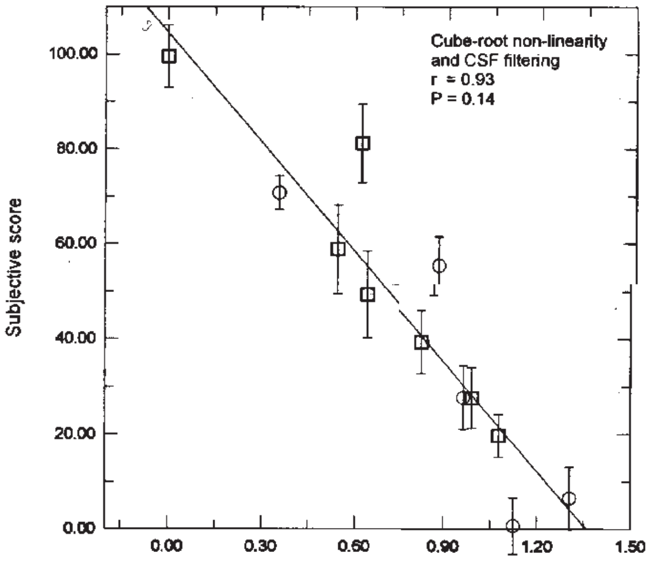
Fig. 4. Continued.

with coloured noise of the bands: [0,1], [1,2], [2,4], [4,8] and [8,16] cycles per degree.

- Experiment II: The task of the observers was assigning a numerical score to the distortion of JPEG compressed images at different bit rates and with non-linear mean square error. To avoid blocking effect distortions, no subblock division was considered: the quantization matrices were applied to the full-image transform. The objective of using this kind of test images is not to evaluate a particular image compression standard, but



(g)

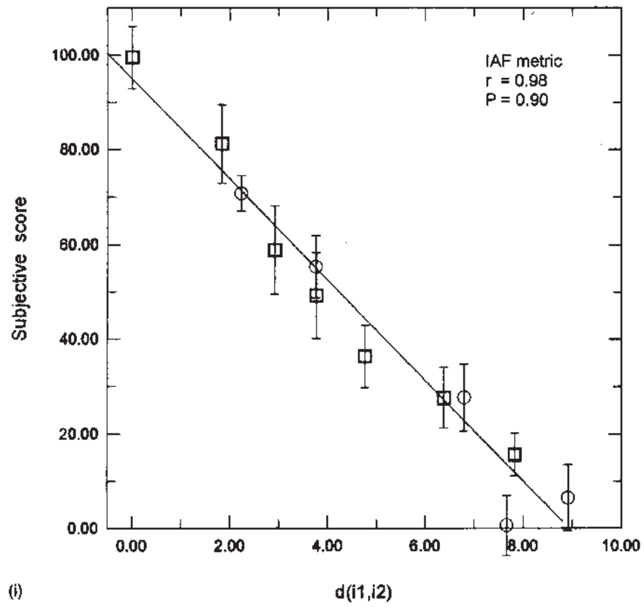


(h)

Fig. 4. Continued.

to generate degraded images with smooth frequency shaped noise of different energies. In experiment II the tests were JPEG-coded versions of natural images to the bit rates: 0.08, 0.13, 0.18, 0.29, 0.38, 0.80 and 5.62 bits/pix.

Examples of the employed test images used in the experiments are shown in Fig. 3. Double stimulus continuous quality scales with adjectival descriptors were used to record the quality judgements [28]. In both cases, the opinions of the observers were normalised to fall in the range [0,100]. Experiments I and II were carried out by 33 and 28 different observers, respectively. 8-bit gray-scale



(i)

Fig. 4. Continued.

images were presented in a 21″ gamma corrected colour monitor (Mitsubishi Diamond Pro 20) subtending 3.8° of visual angle at a viewing distance of 1.5 m. Mean luminance was fixed to be 50 cd m^{-2}. The display was viewed binocularly under fluorescent room lights.

6. Results and discussion

In this section we analyse the performance of our metric and other proposed metrics representing the subjective quality score versus the distance computed with the different algorithms. As experiment I and II are different examples of the same general assessment process, they should follow the same curve when plotted versus algorithms results. In other words, a good metric should give comparable distance values – distance values expressed in the same scale – in spite of the different nature of the evaluated distortions. The free parameters of each model have been fixed as indicated by their respective authors.

Fig. 4 displays the quality assessment of the observers plotted versus the numerical results given by eqns (4)–(12) in experiment I (circles) and experiment II (squares). In Table 1, the goodness of the metrics to match the observers opinion in both experiments is analysed. The correlation coefficient (r), and the value of the χ^2 test for each experiment and metric are given. The χ^2 parameter is the probability that χ^2 is larger or equal to the sum of squared errors ($P(\chi^2 \geq \varepsilon^2)$) In Table 1 the results are given for each isolated experiment and for the whole jointly considered

Table 1
Correlation coefficient, r, and χ^2 parameter, $P(\chi^2 \geq \varepsilon^2)$, for each experiment and metric; each isolated experiment and the whole data are considered; a good metric should be well-behaved in any case

	Experiment I		Experiment II		Both experiments	
	r	P	r	P	r	P
Mean square error (MSE)	–	–	0.93	0.70	0.67	1e − 19
Non-linear MSE	0.13	1e − 18	0.88	0.40	0.68	2e − 17
Gradient-activity masking	0.64	8e − 8	0.98	0.96	0.80	1e − 6
Luminance masking	0.41	3e − 15	0.89	0.45	0.59	2e − 14
Subjective quality factor (SQF)	0.34	1e − 16	0.99	0.99	0.57	6e − 13
Square root integral (SQRI)	0.65	3e − 10	0.96	0.83	0.60	8e − 8
CSF filtering	0.94	0.07	0.97	0.95	0.93	0.13
Non-linearity and CSF filtering	0.91	0.01	0.93	0.72	0.93	0.14
IAF metric	0.97	0.28	0.98	0.96	0.98	0.90

data. A good metric should be well-behaved in any case. P-values below the range [0.01,0.1] indicate that the model does not handle the data [29]. In each figure, the values of r, and P of the fit are given. When the correlation coefficient and the probability of the χ^2 test are marked with an asterisk, (r^*,P^*), it means that the straight line only fits the data of experiment II. In such a case, the metric cannot reproduce the coloured noise data.

Fig. 4 and the confidence data of Table 1 show that some metrics completely fail to predict the results of experiment I. As can be seen in Fig. 3, in both experiments the quality of the distorted test images ranges from very bad to good quality, so if a particular metric cannot deal with the data of a particular experiment, it is due to a lack of generality of its underlying model. The values obtained in the χ^2 confidence measurement may seem too high in some cases, but, as it is well known, these values not only depend on the goodness of the model to match the data, but on the measurement errors too [29]. High variance measurements decreases the sum of squared errors, ε^2, increasing the probability of χ^2 to be larger than ε^2. In this case, high variance is due to the psychophysical nature of the experiments. In this kind of experiments the variance cannot be decreased below certain limits because the observers responses are intrinsically disperse and outliers always occur. This fact may increase the reliability lower limit of P, but the relative ranking remains unchanged. In this case, the experimental data (and their variances) are the same for all the considered models, so they are comparable using the χ^2 parameter.

As we can see in Fig. 4(a), in spite of the clear non-linearity of the results, the MSE fits to a reasonably accuracy the data of the experiment II (squares), but, obviously, it fails to predict the subjective score for the images corrupted by equal energy coloured noise (circles).

Fig. 4(b) shows that if we only include the photoreceptors non-linearities, differences between test images of experiment I are introduced, but the results are not yet valid to handle the subjective opinion in this case.

Fig. 4(c) and 4(d) show the still poor results of the spatial-domain masking based metrics. Although the results of the gradient-activity metric are really good for experiment II, the goodness of the fit for all the data is below the acceptability limit ($P(\chi^2 \geq \varepsilon^2) = 1{\cdot}10^{-6} \ll 0.1$) in spite of the relative good alignment of the data $r = 0.80$. In Fig. 4(e) and 4(f) appear the results of $1/f$ weighted differences in the frequency domain: SQF and SQRI. These metrics achieve very good results when fitting the experiment II data, but they fail again to adjust the results of the experiment I. In the case of Square Root Integral metric, good agreement with subjective quality assessment has been reported in noisy imagery [19], but those results always assume white noise.

The best results are obtained by the CSF-based frequency methods and our IAF metric (Fig. 4(g), 4(h) and 4(i)). These metrics can deal with the results of both experiments giving acceptable fits ($P(\chi^2 \geq \varepsilon^2) > 0.1$) for the whole set of data.

The results show that the equal-energy coloured noise assessment can be reproduced only if the uneven spatial frequency sensitivity of HVS is taken into account (by means of the CSF or the IAF). The SQRI metric do employ the CSF but its effect is perturbed by the $1/f$ factor, overestimating the low frequency contributions to the perceptual error. An accurate prediction of the assessment of frequency dependent or band localised noise is not an academic problem attending to the wide use of the transform domain encoding techniques for still images and video signals. In such techniques the quantization noise is unevenly distributed in the spatial frequency domain according to *statistical* and *perceptual* criteria [6,7,12], so it is very important to know the perceptual effect of selective error allocation.

Fig. 4(g), 4(h) and 4(i) and the quantitative results of the χ^2 test show that there is little benefit from introducing first-stage photoreceptors non-linearities, and that there is a significant improvement in using the fully non-linear model based on bit allocation of HVS.

7. Concluding remarks

In this paper, a new subjective image fidelity metric in the DCT domain has been presented. It is based on an unified fully non-linear and suprathreshold contrast perception model: the Information Allocation Function (IAF) of the visual system. Due to its global empirical nature, this model includes the effects form photoreceptors point-wise non-linearities to post-transform suprathreshold non-linearities.

Exhaustive psychophysical experiments show that the distortion obtained by the IAF-based algorithm is linearly related to the score given by the observers under a variety of noise conditions. The quality of the fit is better in the IAF model than in previously reported metrics. The developed metric can, therefore, be incorporated in the design of compression algorithms as a closer approximation of human assessment of image quality.

Acknowledgements

This work was supported by Generalitat Valenciana/ Diputació de València IVEI scientific research project No. 96/003-035. The authors wish to acknowledge useful discussions with J. García, F. Ferri, J. Albert and P. Capilla.

References

[1] A. Gersho, R.M. Gray, Vector Quantization and Signal Compression, Kluwer, Dordrecht, The Netherlands, 1992.
[2] Y. Linde, A. Buzo, R.M. Gray, An algorithm for vector quantizer design, IEEE Trans. Comm. 28 (1980) 84–95.
[3] J.O. Limb, Distortion criteria of the human viewer, IEEE Trans. on Sys. Man and Cybern. 9 (12) (1979).
[4] H. Marmolin, Subjective MSE measurements, IEEE Trans. on Sys. Man, Cybern. 16 (3) (1986) 486–489.

[5] N. Jayant, J. Johnston, R. Safranek, Signal compression based on models of human perception, in: Proc. IEEE 81 (10) (1993) 1385–1422.
[6] G.K. Wallace, The JPEG still picture compression standard, Comm. ACM. 34 (4) (1991) 31–43.
[7] D. LeGall, MPEG: A video compression standard for multimedia applications, Comm. ACM 34 (4) (1991) 47–58.
[8] ISO/IEC, 13818 draft international standard: Generic coding of moving pictures and associated audio, part 2: Video, 1993.
[9] A.N. Akansu, R.A. Haddad, Multiresolution Signal Decomposition: Transforms, Subbands and Wavelets, Academic Press, San Diego, 1992, pp. 1–7.
[10] B. Macq, Weighted optimum bit allocations to ortogonal transforms for picture coding, IEEE Trans. 10 (1992) 875–883.
[11] K.N. Ngan, H.C. Koh, W.C. Wong, Hybrid image coding scheme incorporating human visual system characteristics, Opt. Eng. 30 (7) (1991) 940–947.
[12] B. Macq, H.Q. Shi, Perceptually weighted vector quantization in the DCT domain, Electr. Lett. 29 (15) (1993) 1382–1384.
[13] J.L. Barron, D.J. Fleet, S.S. Bauchemin, Performance of optical flow techniques, Int. J. Comp. Vis. 12 (1) (1994) 43–77.
[14] S. Haykin, Adaptive Filter Theory, 3rd ed., Prentice Hall, New Jersey, 1996.
[15] I. Pitas, A.N. Venetsanopoulos, Non-linear Digital Filters: Principles and Applications, Ch. 9: Adaptive non-linear filters, Kluwer, Dordrecht, The Netherlands, 1990, pp. 267–311.
[16] N.B. Nill, BR. Bouzas, Objective image quality measure derived from digital image power spectra, Opt. Eng. 32 (4) (1992) 813–825.
[17] LA. Saghri, P.S. Cheatham, A. Habibi, Image quality measure based on a human visual system model, Opt. Eng. 28 (7) (1989) 813–819.
[18] P.J.G. Barthen, Evaluation of subjective image quality with the square-root integral method, JOSA A. 7 (10) (1990) 2024–2031.
[19] P.J.G. Barthen, Evaluation of the effect of noise on subjective image quality, in: Proc. of the SPIE 1453, Human Vision, Visual Processing and Digital Display II, 1991.
[20] D.J. Granrath, The role of human visual models in image processing, Proc. IEEE 69 (5) (1981) 552–561.
[21] D.R. Fuhrmann, J.A. Baro, J.R. Cox, Experimental evaluation of psychopysical distortion metrics for JPEG-encoded images, Proc. of the SPIE 1913 (1993) 179–190.
[22] W.K. Pratt, Digital Image Processing, John Wiley, New York, 1992.
[23] S. Daly, Application of a noise-adaptive Contrast Sensitivity Function to image data compression, Opt. Eng. 29 (8) (1990) 977–987.
[24] D.H. Kelly, Receptive-field-like functions inferred from large-area psychophysical measurements, Vision Res. 25 (12) (1985) 1895–1900.
[25] G.E. Legge, A power law for contrast discrimination, Vision Res. 21 (1981) 457–467.
[26] J. Malo, A.M. Pons, J.M. Artigas, Bit allocation of the human visual system inferred from contrast incremental thresholds of sinusoidal gratings, Perception 24 (Supl.) (1995) p. 86.
[27] J. Malo, A.M. Pons, J.M. Artigas, Bit allocation algorithm for codebook design in vector quantization fully based on human visual system non-linearities for suprathreshold contrasts, Electr. Lett. 24 (1995) 1229–1231.
[28] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, D. Pearson, Measurement of scene-dependent quality variations in digitally coded television pictures, IEE Proc. Vis. Image and Signal Proc. 142 (3) (1995) 149–154.
[29] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, 2nd ed., Ch. 15: Modelling of data, Cambridge University Press, Cambridge, 1992, pp. 659–666.

A.2 Characterization of the HVS threshold performance by a weighting function in the Gabor domain

Journal of Modern Optics. Vol. 44, N. 1, pp. 127–148 (1997)

JOURNAL OF MODERN OPTICS, 1997, VOL. 44, NO. 1, 127–148

Characterization of the human visual system threshold performance by a weighting function in the Gabor domain

J. MALO, A. M. PONS, A. FELIPE and J. M. ARTIGAS
Departamento Interuniversitario de Óptica, Facultat de Física, Universitat de València, C/Dr. Moliner 50, 46100 Burjassot, València, Spain. e-mail: Jesus.Malo@uv.es

(Received 2 February 1996; revision received 28 May 1996)

Abstract. As evidenced by many physiological and psychophysical reports, the receptive fields of the first-stage set of mechanisms of the visual process fit to two-dimensional (2D) compactly supported harmonic functions. The application of this set of band-pass filter functions to the input signal implies that the visual system carries out some kind of conjoint space/spatial frequency transform. Assuming that a conjoint transform is carried out, we present in this paper a new characterization of the visual system performance by means of a weighting function in the conjoint domain. We have called this weighting function (in the particular case of the Gabor transform) the Gabor stimuli Sensitivity Function (GSF) by analogy with the usually employed weighting function in the Fourier domain: the Contrast Sensitivity Function (CSF). An analytic procedure to obtain this conjoint weighting function from the psychophysical measurements of the CSF is derived. The accuracy of the procedure is proved showing the equivalence between some experimental Fourier weighting functions (2D CSFs), and the corresponding GSFs. The main advantage of this new characterization is that a weighting function in a space/spatial frequency domain can account for spatially variant behaviour, which cannot be included in a unique CSF, so a single GSF would be needed to include extra-foveal and large eccentricity behaviour. One example is given of how non-homogeneous systems can be easily characterized in this way.

1. Introduction

Among the main research objectives in the field of spatial vision, there is the implementation of algorithms for visual information encoding which, in accordance with physiology, could reduce the signal volume for the next processing stage. The strong correlation between luminance values in neighbouring points of the natural images [1] points to performing the information processing in a domain different from the spatial one [2, 3]. In this way, one of the tasks of the visual system must be to transform the input images, by projecting them over a more appropriate domain for their analysis.

Since the publication of the paper by Campbell and Green on contrast sensitivity to sinusoidal gratings [4], the visual system has been commonly characterized by accepting that it performs a Fourier transform (FT) of the input signal and, subsequently, the FT's coefficients are selectively attenuated by a weighting function defined in the spatial frequency domain: the CSF. In the 1980s several physiological studies [5–7] showed that the receptive fields of the cortical cells resemble compactly supported harmonic functions. At the same time, the interest on the intrinsic two-dimensionality of the visual process led Daugman [8, 9] to question the frequency/orientation separability of the spatial vision mechanisms. In these pioneering works [8, 9], limited band-pass Gabor filter functions are proposed to characterize the receptive field and the spectral response of the mechanisms. Detailed 2D-oriented psychophysical measurements carried out by Daugman [10] and subsequently confirmed by Harvey and Doan [11] prove that in the initial stages of the visual process, the input signal is analysed by a set of filters simultaneously localized in the spatial and the frequency domains.

Since then, the papers of Daugman [12, 13], Porat and Zeevi [14], Field [3, 15], and Watson [16–18], among others, develop models where the analysing transform is no longer a Fourier transform, but some kind of conjoint space/spatial frequency transform more consistent with the experimental evidences. All these models describe different linear conjoint transforms (CT) applied by the system to place the signal into a more suitable domain for its analysis, but, in fact, these transforms can be considered as simple changes of domain, giving no clue about the real information process carried out by the system.

In this work we characterize these information reduction processes by means of a weighting function defined in the domain of the conjoint transform (CT). This weighting function acts on the coefficients of the CT just as the CSF acts on the coefficients of the FT. So, assuming that the system performs a generic linear CT as described in the literature, our contribution is focused in the following points:

- The description of the information reduction processes by a weighting function in a 4D space–frequency conjoint domain.
- To propose an analytic procedure to obtain the values of this 4D weighting function from the psychophysical data of the equivalent weighting function in the 2D Fourier domain (the 2D CSF).
- To test the proposed algorithm with experimental 2D CSF data.
- To emphasize, with a how-to example, the potentialities of this characterization to include in a single function a spatially variant behaviour, what is not possible in a Fourier characterization.

Finally, we want to point out that even though we have assumed the generic CT to be a Gabor transform (GT), it is not our objective to elucidate whether the Gabor transform is in fact the linear transform carried out by the visual system. This matter is still in discussion [19]. Anyway, the mathematical reasonings are completely generalizable to any other kind of linear conjoint transform, by changing GT for CT in the expressions.

2. Space–frequency analysis as an alternative to Fourier frequency analysis in the spatial vision field

We shall begin by remembering the basic ideas about space–frequency analysis and its advantages over Fourier frequency analysis.

In the contrast range where system behaviour is approximately linear, it is accepted that the system performs a transformation of the topographical signal $f(\mathbf{x})$ in such a way that $f(\mathbf{x})$ is then written as a linear combination (either continuous or discrete) of a set of base functions $\{g(\mathbf{x}, \mathbf{p})\}$ which are dependent on a parameter (or parameters) $\mathbf{p}$. That is:

$$f(\mathbf{x}) = \int_P C(\mathbf{p}) g(\mathbf{x}, \mathbf{p})\, \mathrm{d}\mathbf{p}, \tag{1}$$

where the coefficients $C(\mathbf{p})$ indicate the participation degree of each $g(\mathbf{x}, \mathbf{p})$ function in the signal $f(\mathbf{x})$ or, in other words, to what extent the signal contains the characteristics of each base function.

The different approximations to space–frequency analysis (wavelet analysis [20–24], Gabor analysis [25] ...) have a common property which makes them different from Fourier analysis: the simultaneous localization of the base functions $g(\mathbf{x}, \mathbf{p})$ in the space and frequency domains. Thus, the coefficients $C(\mathbf{p})$ give information about $f(\mathbf{x})$ in both domains.

Within the space–frequency analysis, the method of obtaining the base functions and their analytical expressions constitute the main difference between the various approaches. In wavelet analysis, the set of base functions is obtained from one single generating function by expansions, compressions, translations and rotations [20, 23]. The elegance of this method has a drawback: the nonexistence of an explicit analytical expression for calculating the generating function. The kind of functions that can be considered as generating ones are limited by some theorems [22].

In the Gabor analysis the base is formed by harmonic functions of frequency $\mathbf{k}_0$ which are modulated by a Gaussian window centred in any point $\mathbf{x}_0$ of the spatial domain [25]:

$$g(\mathbf{x}, \mathbf{p}) = g(\mathbf{x}, \mathbf{x}_0, \mathbf{k}_0) = C \exp\left[-\left(\frac{\mathbf{x} - \mathbf{x}_0}{a}\right)^2\right] \exp -\mathrm{i}\mathbf{k}_0\mathbf{x}. \tag{2}$$

Here C is just a constant to make $|g(\mathbf{x}, \mathbf{x}_0, \mathbf{k}_0)| = 1$. The Gaussian envelope, thanks to the simplicity of its analytic expression, allows a simultaneous and intuitive control of the shape of the function and its spectrum by adjusting the parameters of its formula [12]. Moreover, the Gabor function displays two fundamental properties: it makes possible parallel treatments in both frequency and space domains [26], and it minimizes the product of the widths in the space and frequency domains [12, 27]; thereby, the information supplied by each coefficient is maximal.

The advantages of space–frequency analysis are basically three:

(1) The image analysis from the coefficients of either its wavelet transform (WT) or Gabor transform (GT) facilitates the high-level process task. This is due to the fact that it is possible to assign a spatial and frequency significance to each coefficient (of either WT or GT).
(2) In either a hardware or neural implementation of the space–frequency transformations, the number of connections between pixels is minimized. This fact is due to the control of the spatial width as a function of the central frequency of each base function, which minimizes the image portion employed in calculating each coefficient [22]. On the contrary, in an FT, every image point must be employed in calculating each coefficient.
(3) Strong compression factors of natural images are obtained through the space–frequency representations [3, 13, 28–30]. The particular statistic of

these images leads to the fact that very few coefficients reach significant values, dramatically reducing the transformed function volume.

The introduction of this kind of analysis in the spatial vision field is due to two reasons; on the one hand, to the recognition of its intrinsic advantages over the Fourier analysis [3, 9, 13] and, on the other hand, to the physiological [5–7] and psychophysical [10, 11] evidences which indicate that the Gabor functions are the best fitting for receptive fields and frequency tuning mechanisms.

Criteria mentioned before have favoured the employment of this representation as a model for spatial vision [4, 31, 32] and for texture perception [31, 33] as well as the employment of Gabor stimuli in psychophysical experiments to study the interaction between the different mechanisms of detection in pattern recognition [15], or the induction and masking effects [34].

3. Gabor stimuli Sensitivity Function (GSF)

Assuming that the visual system has performed a GT of the input image, there exist two possibilities for explaining the information reduction which takes place during the processing. One possibility consists in proposing that the transformation is achieved by means of an incomplete basis and, therefore, when the image is decomposed with regard to such a basis there is a loss of information. The problem with this option is that the choice of the basis is decisive to obtain a suitable characterization, but establishing a criterion to adjust the parameters of each base function is not an easy task. The other possibility is to assume that the GT achieved is complete (or quasi-complete [26]), so the loss of information is placed in a selective weighting process after the transformation. This is the approach we have followed. In this case, the choice of the base becomes less critical, since the problem is now the calculation of the weights. This calculation, as we will see, can be formulated independently of the base selected.

We have called the Gabor stimuli Sensitivity Function (GSF) the weighting function acting on the GT coefficients, in the visual system case. This model of the encoding and initial filtering can be summarized in the block diagram shown in figure 1.

Linear behaviour of the system is required, as in the Fourier [35] case, since the GT and the weighting process after it are linear. Therefore, the model validity is, at first, limited to the low-contrast range. Nevertheless, the characterization of a system in the Gabor domain is less restrictive than the Fourier one, since system homogeneity is not an indispensable requirement here, as it is in the Fourier formalism due to the global character of the FT.

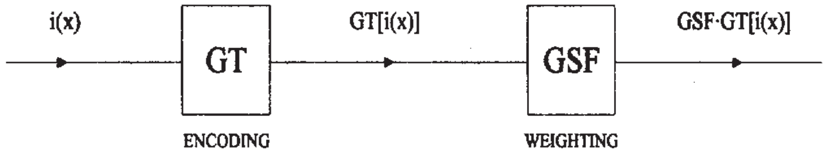


Figure 1. Block diagram of the transformations of the input signal (a still achromatic image) to obtain its distorted representation at cortex level (as a set of neural impulses of cells with Gabor-like receptive fields). The weighting function GSF includes both optical and neural degradation processes.

4. Relation between the weighting functions defined in the Fourier and Gabor domains

To establish the relation between both functions, we assume the equivalence of both system characterizations. That is, the same degraded image should be obtained if it were reconstructed from the output coefficients of either case. This is shown schematically in figure 2 and represented mathematically as

$$\begin{aligned} i'_{\mathrm{CSF}}(\mathbf{x}) &= \mathrm{FT}^{-1}[\mathrm{CSF}(\mathbf{k})\cdot \mathrm{FT}[i(\mathbf{x})](\mathbf{k})](\mathbf{x}) = i'_{\mathrm{GSF}}(\mathbf{x}) \\ &= \mathrm{GT}^{-1}[\mathrm{GSF}(\mathbf{x},\mathbf{k})\cdot \mathrm{GT}[i(\mathbf{x})](\mathbf{x},\mathbf{k})](\mathbf{x}). \end{aligned} \quad (3)$$

Note that this expression (the relation between the Fourier and the conjoint weighting functions) is independent of the conjoint basis, so it can be applied to any general CT. The main restriction of equation (3) comes from the fact that it relates a 4D object (the conjoint weighting function) to a 2D object (the CSF). The calculation of the GSF from the CSF, using equation (3), implies the spatial independence of the GSF coefficients. In fact, as the CSF characterization requires homogeneity (spatial invariance) of the system, it does not include the possibility that the system response to a particular spatial frequency varies spatially. Nevertheless, taking into account that spatial dependence is the advantage of the GSF, as this characterization assigns several coefficients to each frequency from a stimulus of that frequency situated at different spatial position, we should include the spatial dependence effects. This can be done introducing a different attenuation over the coefficients corresponding to the same frequency but with different spatial positions in order to get a spatially variant point spread function of the system.

For all these reasons, it is apparent that equation (3) will only provide a first approximation to the GSF. The GSF obtained from equation (3) will determine an attenuation dependent on the frequency, but uniform at a certain frequency for each spatial position. As well as the CSF used in the calculation, and its corresponding point spread function (having a validity limited to the foveal region where the response to the unit stimuli remains constant), this first approximation to the GSF is only valid if it is locally applied. Although equation (3) introduces

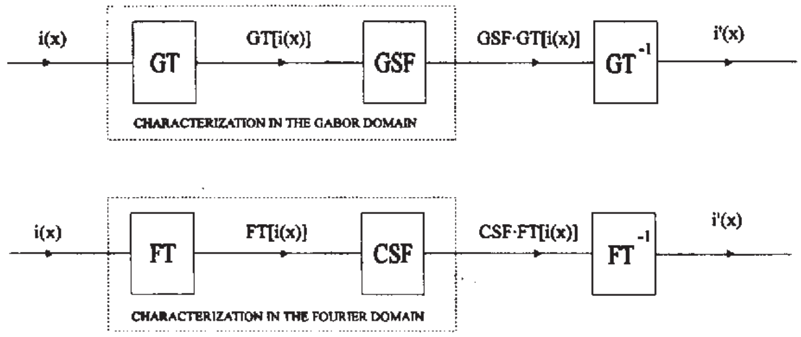


Figure 2. Obtention of a topographic signal from the encoded signal in both Gabor and Fourier representations applying the appropriate inverse transform. If the characterized system is the same, the result of both weighting processes must be the same.

local validity, this restriction can be overcome introducing a selective spatial attenuation for the coefficients of each frequency.

Having presented equation (3) and its restrictions, we shall mention the explicit relations between the CSF and the GSF. Making $i(\mathbf{x})$ in equation (3) equal to a base function $\exp(i\mathbf{k}_0\mathbf{x})$ or $g(\mathbf{x},\mathbf{x}_0,\mathbf{k}_0)$, we can see that

$$\mathrm{GSF}(\mathbf{x}_0,\mathbf{k}_0) = \mathrm{GT}[\mathrm{FT}^{-1}[\mathrm{CSF}(\mathbf{k})\cdot \mathrm{FT}[g(\mathbf{x},\mathbf{x}_0,\mathbf{k}_0)](\mathbf{k})](\mathbf{x})](\mathbf{x}_0,\mathbf{k}_0), \tag{4a}$$

$$\mathrm{CSF}(\mathbf{k}_0) = \mathrm{FT}[\mathrm{GT}^{-1}[\mathrm{GSF}(\mathbf{x},\mathbf{k})\cdot \mathrm{GT}[\exp(i\mathbf{k}_0\mathbf{x})](\mathbf{x},\mathbf{k})](\mathbf{x})](\mathbf{k}_0), \tag{4b}$$

if it is normalized. The GSF of equation (4 *a*), as mentioned above, does not involve system inhomogeneities, and it will have a local validity restricted to the region where the CSF, from which it is calculated, is valid. This problem can be solved by introducing spatial selectivity *a posteriori*.

If a general GSF, including the system inhomogeneities, is employed to calculate the CSF from equation (4 *b*), those inhomogeneities will be averaged in the calculation. Since a filter defined in a 4D domain is projected into a 2D domain, thereby all the inhomogeneities present in the 2D removed domain are averaged. This fact implies a foveal CSF undervaluation and an extra-foveal CSF overvaluation. To avoid this average over very different areas, we can calculate local CSFs from the spread function of the system for a certain area, using the equation

$$\mathrm{CSF}_{\mathbf{x}_0}(\mathbf{k}) = \mathrm{FT}[\mathrm{GT}^{-1}[\mathrm{GSF}(\mathbf{x},\mathbf{k})\cdot \mathrm{GT}[\delta(\mathbf{x}-\mathbf{x}_0)](\mathbf{x},\mathbf{k})](\mathbf{x})](\mathbf{k}). \tag{5}$$

In short, equations (4 *a*) and (5) allow us to make the GSF calculation (which does not include spatial attenuation) and the CSF calculation (with validity in the proximity of $\mathbf{x}_0$) from knowledge of the CSF and GSF, respectively.

As we can see, equations (4 *a*) and (5) are independent of the basis $\{g(\mathbf{x},\mathbf{x}',\mathbf{k})\}$ chosen. This fact provides a certain freedom in basis selection, and then the characteristics of the visual system detection mechanisms [5–7] can be considered to a greater or lesser degree. Thus, in addition to logical dependence on the CSF, the concrete values of the GSF will be dependent on the base chosen for performing the transform.

This global-to-local drawback inherent in any 2D to 4D projection would be present in any other conjoint representation, but, as stated above, the expressions (4 *a*) and (5) are generalizable to any other linear CT, such as wavelet transforms, just by changing GT by the proper CT and making use of the proper basis (wavelets or any kind of windowed sinusoids), so the problem is equally overcome.

5. Determination of the 4D GSF from a 2D CSF

Three things are required for calculating the GSF according to equation (4 *a*):

- the Gabor's channels model: a set of base functions over which the decomposition of the signal should be carried out,
- the algorithm for calculating the GT with the chosen basis,
- experimental data about the weights in the Fourier domain.

The model of Gabor's channels employed will have to include the following properties of the visual system channels:

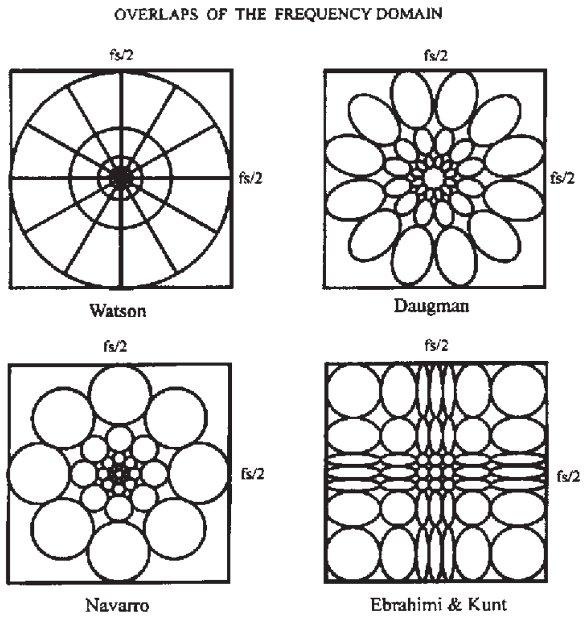


Figure 3. Different models to overlap the frequency domain with the spectra of the base functions (f_s is the sampling frequency).

- an approximately constant width in orientation (about 30°),
- a frequency width of about 1 octave,
- four or five frequency tuning mechanisms with sensitivity maximums separated by octaves.

A rigorous reproduction of these characteristics will lead us to propose an overlap of the frequency space similar to those proposed by Daugman [13], Watson [17, 18] and Navarro and Tabernero [26] (see figure 3).

We have chosen the separable base functions of Ebrahimi and Kunt [28] because, although they do not strictly fit the physiological data, they reflect their essential properties: the overlap is more compact in the low-frequency zone, the frequency width increases with the frequency, and the separation between central frequencies is also about 1 octave. Moreover, this overlap implies a lower computational cost than other algorithms, since the calculation of GT and GT^{-1} is reduced to a matrix product.

In any case, in this paper the algorithm for the GT calculation and the basis model employed are only tools and, therefore, their properties are not essential aspects of our discussion, since our main aim is to obtain the GSF weights in the Gabor domain. To do this we merely use a certain base and algorithm to apply equations (4 *a*) and (5). Consequently, the choice of a certain model, perhaps not the most appropriate, does not disminish the validity of the reasoning.

Using a separable basis, the functions $g_{kl}(x, y)$, defined in a 2D discrete domain of $N \times N$ points, can be written as an outer product of the functions $g_k(x)$ and $g_l(y)$ defined in a 1D domain of N points. That is:

$$[g_{kl}(x,y)] = [g_l(y)]^{[g_k(x)]}, \tag{6}$$

and, therefore, the discrete linear combination analogous to equation (1),

$$f(x) = \sum_{k,l} C_{kl} g_{kl}(x,y), \tag{7}$$

has a matrix expression

$$\mathbf{f} = \mathbf{G}_1 \mathbf{C} \mathbf{G}_2, \tag{8}$$

where $\mathbf{f}$ is the matrix of image points, the columns (rows) of the matrix $\mathbf{G}_1$ ($\mathbf{G}_2$) have as elements the 1D functions g_i, and the matrix $\mathbf{C}$ has as elements the coefficients that control the contribution of each $g_{ij}(\mathbf{x})$ in $f(\mathbf{x})$.

To obtain the GT of a function $\mathbf{f}$ with respect to a separable base implies finding the matrix $\mathbf{C}$, so that the mean square error (mse) is minimized:

$$\text{mse} = \|\mathbf{f} - \mathbf{G}_1 \mathbf{C} \mathbf{G}_2\|^2. \tag{9}$$

This kind of minimum square problem has the solution [36]

$$\mathbf{C} = \mathbf{V}_1 (\mathbf{S}_1^{\mathrm{T}})^{-1} \mathbf{U}_1^{\mathrm{T}} \mathbf{f} \mathbf{V}_2 (\mathbf{S}_2^{\mathrm{T}})^{-1} \mathbf{U}_2^{\mathrm{T}}, \tag{10}$$

where the matrices $\mathbf{U}_i, \mathbf{S}_i$ and $\mathbf{V}_i$ arise from the decomposition in singular values of the matrices $\mathbf{G}_i$. Thus, once a basis is selected, and the matrices $\mathbf{G}_i, \mathbf{U}_i, \mathbf{S}_i$ and $\mathbf{V}_i$ are calculated, the calculation of GT and GT^{-1} is reduced to computing the matrix products (10) and (8), respectively.

In this representation, the GSF filter will be given by a matrix with the same dimensions as $\mathbf{C}$, so that every coefficient GSF_{ij} acts on every coefficient C_{ij} of the GT of the image $\mathbf{f}$, in such a way that the filtered image $\mathbf{f}'$ will be

$$\mathbf{f}' = \mathbf{G}_1 \cdot \mathbf{GSF}^* \mathbf{C} \cdot \mathbf{G}_2, \tag{11}$$

where the dot symbol $\cdot$ represents the usual matrix product and $*$ is the element by element product of the GSF matrix and $\mathbf{C}$:

$$(\mathbf{GSF}^* \mathbf{C})_{ij} = \mathrm{GSF}_{ij} C_{ij}. \tag{12}$$

The space–frequency meaning of every coefficient C_{ij} or GSF_{ij} depends on how the 1D functions g_i are organized in the matrices $\mathbf{G}_i$. In our case, the base functions $g_{kl}(\mathbf{x})$ cover the frequency space as in the Ebrahimi and Kunt [28] model. The covering of the spatial domain by functions of a given frequency (not well clarified in [28]) is performed by imposing a constant overlapping, independent of the frequency, of the contiguous functions. In this manner, the low-frequency functions (of narrow spectrum and, therefore, large spatial width) are very separate in the spatial domain, so that it is covered by a few of these functions, whereas the high-frequency functions (of wide spectrum and, therefore, very spatially localized) are very close, requiring a greater number of functions by unit of area to cover the spatial domain.

Our model uses a basis with seven central frequencies in each direction: $3f_s/8, f_s/6, f_s/18, 0, -f_s/18, -f_s/6, -3f_s/8$ (f_s being the sampling frequency), i.e.

49 different frequencies and as many spatial positions as necessary to satisfy the imposed overlapping conditions. To determine an overlapping value, the criterion was the achievement of a perfect reconstruction in developing the GT and GT^{-1}. (As shown in figure 1, our model consists in assuming a *complete* encoding followed by a filtering process).

In our case, an overlapping of 85% was imposed in order to get reconstructions with a negligible error. It is apparent that the selection of a different covering of the frequency domain could perhaps have diminished the percentage of overlapping needed, but, as we said before, our goal is to use the Gabor channels model as a tool and not to optimize it.

The matrix $\mathbf{G}_1$ (or $\mathbf{G}_2$) is constituted by submatrices of columns (rows) corresponding to the different frequencies, from the most negative ones on the left (above) to the most positive ones on the right (below). Inside each of these submatrices, every column (row) corresponds with a different spatial position, being also arranged in a crescent from left to right (from top to bottom). This particular distribution of the 1D functions in the matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ implies a determinated space–frequency meaning for every coefficient of the transform $\mathbf{C}$ or of the filter GSF (see one example of this in figure 4).

The beneficial effects of all the advantages of the GSF are obtained if the coefficients are calculated in the whole 4D domain (2D spatial domain and 2D frequency domain). We apply equation (4 *a*) using the data of CSFs evaluated in the whole 2D domain in such a way that they include the asymmetries in orientation of the system under study [37]. In particular, the experimental data supporting this paper are the CSFs of two observers [37]: JM (emmetropic) and RZ (astigmatic). Figure 5 shows the filter functions for both these observers.

The results of applying equation (4 *a*) to the data of JM, for every function $g_{kl}(\mathbf{x})$, is shown in figure 6 (*a*). As we can see, the weights are organized as a band-pass function, with maximum sensitivity for the functions with central frequency of 1·67 cpd ($f_s = 60$ cpd). The frequency resolution in the Gabor representation is lower than in the Fourier representation: Gabor functions have a large band and, therefore, the asymmetries of the Fourier weighting function are averaged in the Gabor representation. As we will see later, this fact does not mean that the information about the astigmatism is lost.

For a certain spatial frequency, no significant difference is obtained for the diverse spatial coefficients, as was expected from the analysis of the restrictions of equation (4 *a*), since the CSF assumes a spatial-invariant behaviour. Therefore, in the GSF there is an attenuation varying for the different submatrices corresponding to the different frequencies; nevertheless, a constant attenuation is found for all the coefficients of a given frequency corresponding to every spatial position of the function of that frequency.

A small difference is obtained for certain coefficients, especially for those corresponding with the functions more localized in frequency (and hence much more extended in space), this effect being stronger for the most external positions. This systematic deviation is caused by the finite character of the spatial domain considered, in such a way that a sort of spectral leakage [38] occurs. The function with larger spatial width placed on the border of the domain has non-zero values beyond the limits of the considered domain. This produces a distortion in the result of equation (4 *a*) with regard to other positions, in which the area where the function has a significant value is completely contained in the domain considered.

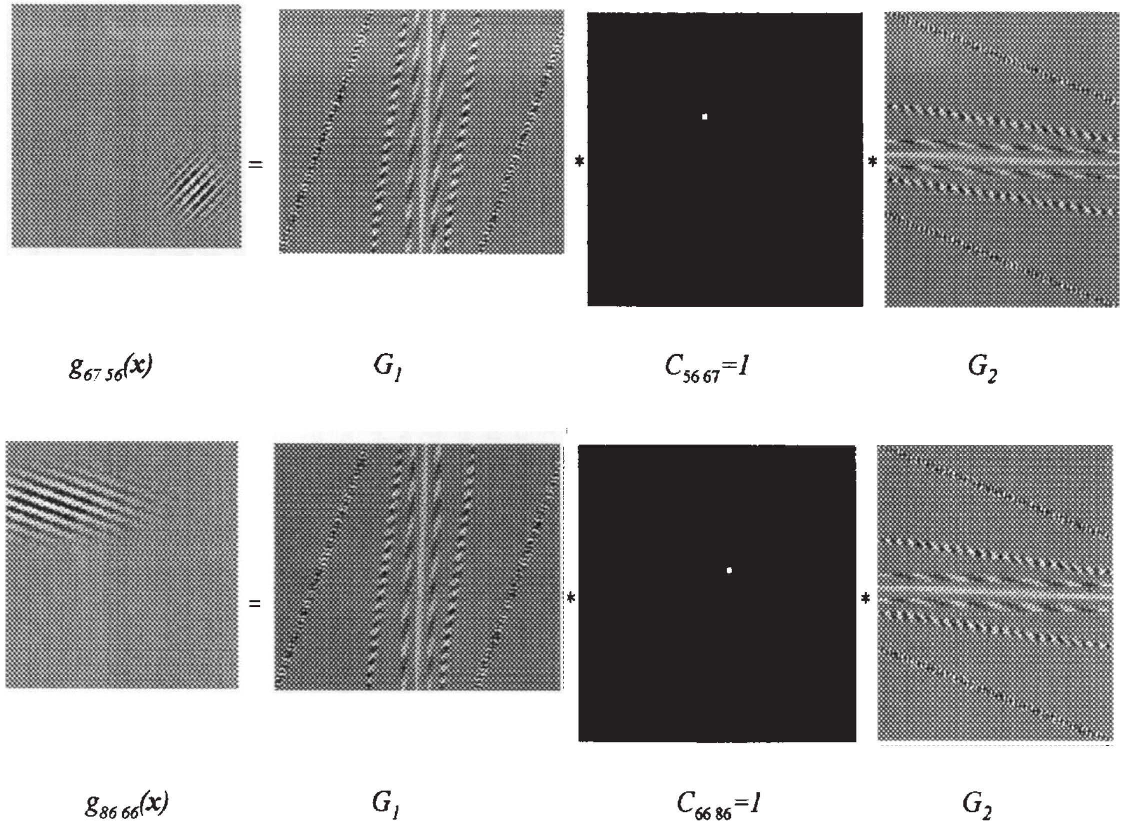
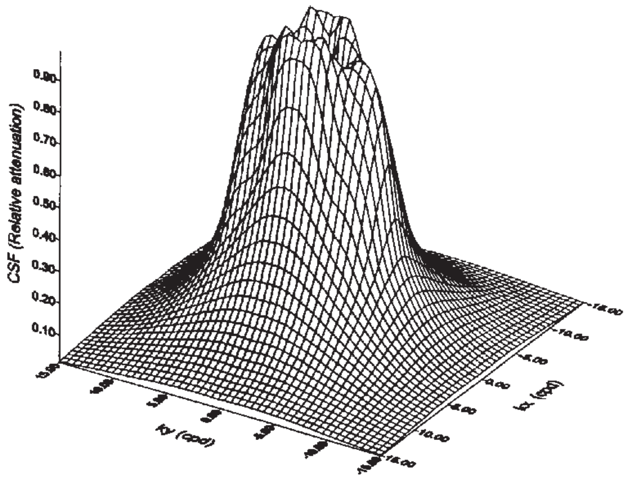
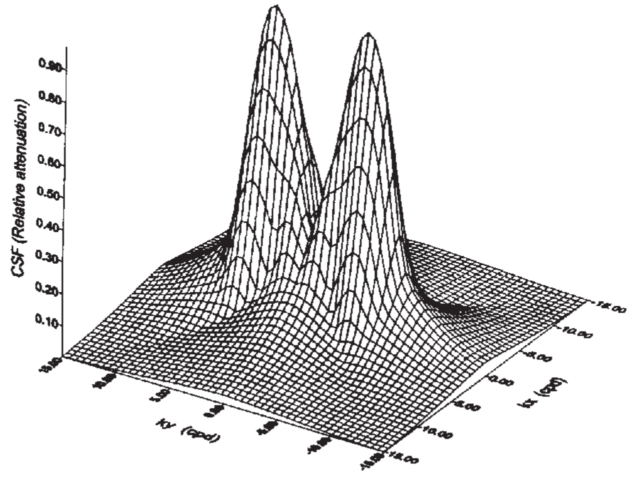


Figure 4. Two examples of the meaning of the GT coefficients: each element of the **C** matrix selects one column of $\mathbf{G}_1$ and one row of $\mathbf{G}_2$ to generate the corresponding base function. In these examples we can see how the coefficients $C_{56\,67}$ and $C_{66\,86}$ (any other elements in the matrix are zero) give rise to the functions $g_{67\,56}(\mathbf{x})$ and $g_{86\,66}(\mathbf{x})$, respectively.



(*a*)

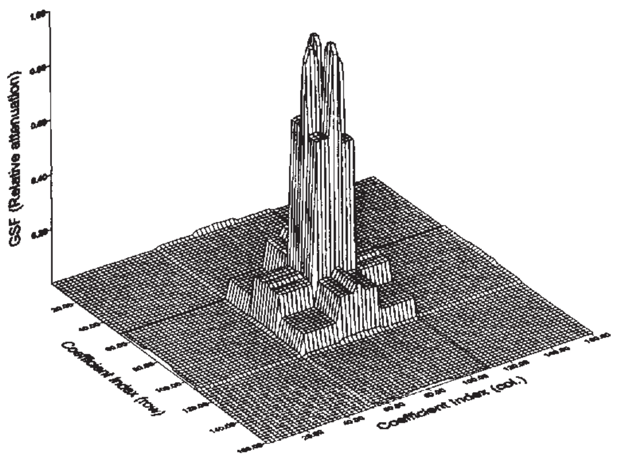


(*b*)

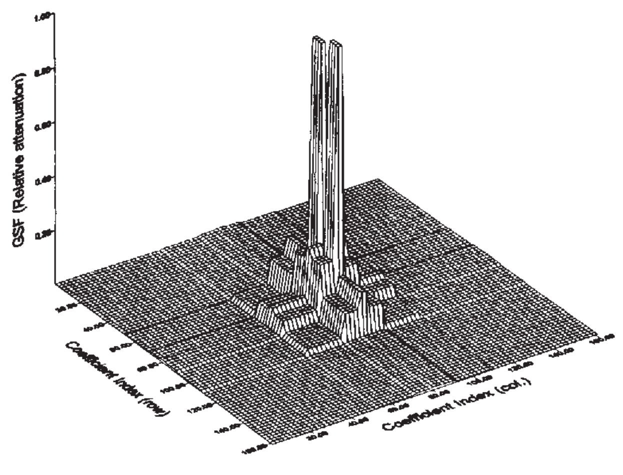
Figure 5. Experimental 2D CSFs of the observers. (*a*) Observer JM (emmetropic), (*b*) Observer RZ (astigmatic -3 dp 0°). Note how the revolution symmetry of the CSF breaks down in this case, favouring vertical frequencies. For details on the measurement of these functions, see [37].

These deviations are only originated by the calculation in a spatially limited domain, and therefore they are not connected at all with the system characteristics. The real value of the attenuation for these stimuli will be given by the results obtained for functions completely contained in the spatial domain.

According to all these considerations, in order to obtain the GSF from equation (4 *a*), it is necessary to calculate only the GSF_{ij} coefficients for the $g_{ij}(\mathbf{x})$ functions



(*a*)



(*b*)

Figure 6. GSFs of the observers obtained from the CSFs of figure 5 by means of equation (4 *a*). The space–frequency meaning of the different coefficients is explained in the text. (*a*) Fully calculated GSF of the observer JM (emmetropic). Note the unexpected deviations of the coefficients of the border of each submatrix (the more external spatial coefficients). These small deviations are due to the finite length of the considered spatial domain. Note that the attenuation depends basically on the magnitude of **k**, but not on the orientation of the frequency vector. (*b*) Sparse calculated GSF of the observer RZ (astigmatic) avoiding the border errors (see text). It is remarkable to see the strong differences in the attenuation of the central submatrices caused by astigmatism.

placed in the centre of the spatial domain considered, avoiding, in this way, border errors and also saving a great amount of calculation. The value calculated for the central coefficient of every submatrix is extended to the rest of the submatrix coefficients to get the final GSF. Using this procedure, we obtain the filter function for the astigmatic observer (figure 6 (*b*)).

6. Equivalence between the spatially invariant GSF and the CSF in the foveal region

To demonstrate the equivalence between the computed GSFs and the CSFs from which they have been calculated, we simulate the spread function of the system and the complex image filtering by means of the GSF. Afterwards, we compare such simulations with other ones from the experimentally measured CSFs. The system spread function calculation and the complex image filtering from the experimental data of a 2D CSF were carried out according to the method presented in previous work [37] The filtered image of an object test from the GSF data was calculated using equation (11), where the matrix **C** corresponds with the GT coefficients of the object test.

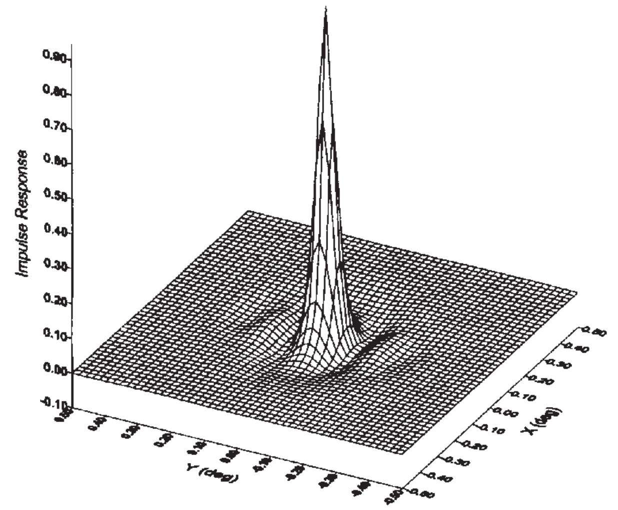
Figures 7 and 8 show the results of the point spread functions (PSFs) for the two observers obtained from their GSFs (figures 7 (*b*) and 8 (*b*)), and compared with the analogous results obtained from the experimental measures of the CSFs (figures 7 (*a*)) and 8 (*a*)). As we can see, the width at half height of the PSF is adequately reproduced for both observers, as well as the PSF deformation along the x axis in the astigmatic observer's case. Although the astigmatism is not clearly manifested in the shape of the GSF because of its lower frequency resolution, this representation keeps the information about the system asymmetries in orientation.

The agreement between both characterizations is also observed in the simulations of complex object processing (figures 9 and 10). Figure 9 shows the processing of a natural image, whereas in figure 10 we employed a circular test to manifest the asymmetries in orientation. The general degradation of the original image (figure 9 (*a*)), when weighted by the GSF of the observer JM (figure 9 (*b*)) and by the CSF of the same observer (figure 9 (*c*)), is equivalent: the difference in mean square error is only about 12%. Figures 9 (*d*) and (*e*) show analogous results for the observer RZ. The difference between both results in this case is 18%. As can be observed in figure 10, the GSF reproduces well the astigmatic behaviour of the observer RZ. The horizontal borders appear much more sharply than the vertical ones, both when the image has been weighted by the GSF (figure 10 (*b*)) and when Fourier simulation is used. For the emmetropic observer (JM) we can see that the degradation appearing in the test is isotropic (figures 10 (*d*) and (*e*)).

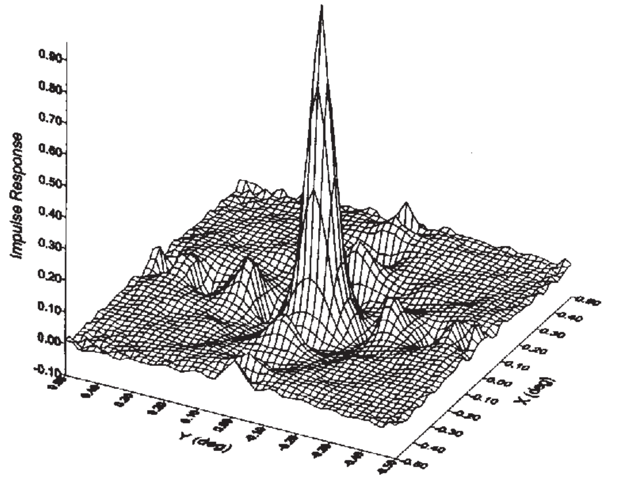
The results demonstrate the accuracy of the assumption on which equation (4 *a*) is based, and the validity of the particular results for the GSFs of the observers JM and RZ, obtained from their CSFs in the foveal region.

7. Spatial non-homogeneities introduced in the GSF: characterization of spatially variant systems

In this section we demonstrate that it is possible to introduce spatial non-homogeneities in the GSF, so that the characterized system will have a spatially variant behaviour. To do this, we infer a simulated observer from the data of the real observers in such a way that the synthesized observer has a different



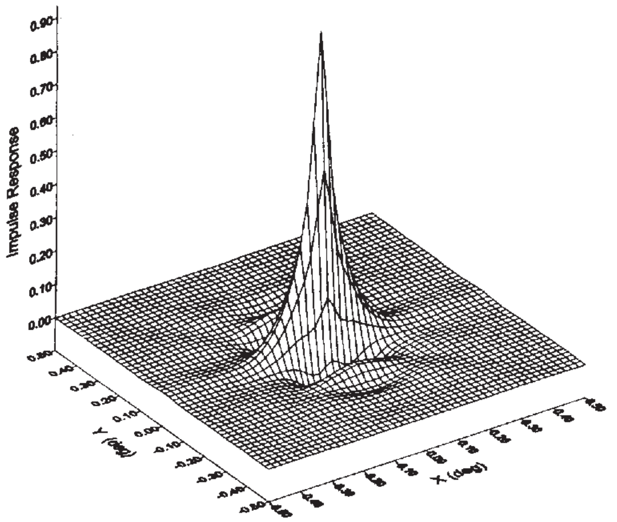
(*a*)



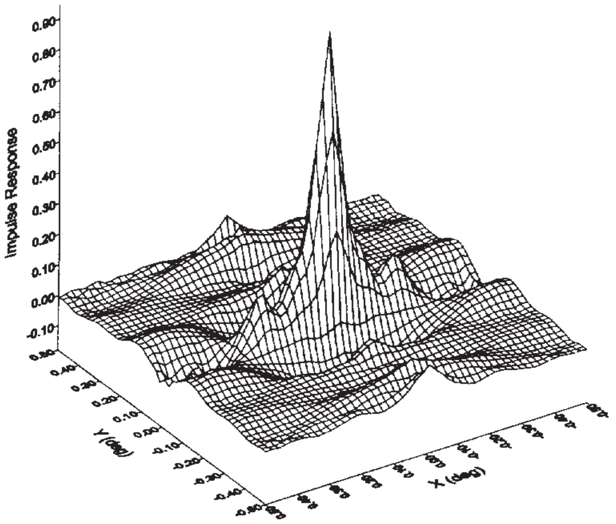
(*b*)

Figure 7. Point spread functions of the emmetropic observer obtained from (*a*) the Fourier characterization and (*b*) the Gabor characterization.

behaviour, emmetropic or astigmatic, in different spatial zones. That is, we have synthesized a non-homogeneous observer. It is qualitatively shown in this simulation how the non-homogeneities of the system can be introduced to characterize the vision in large fields. A realistic model of this type of perception would necessarily include, in a systematic form, data about the system response to different eccentricities [39, 40], which would simply imply an increment in the



(*a*)



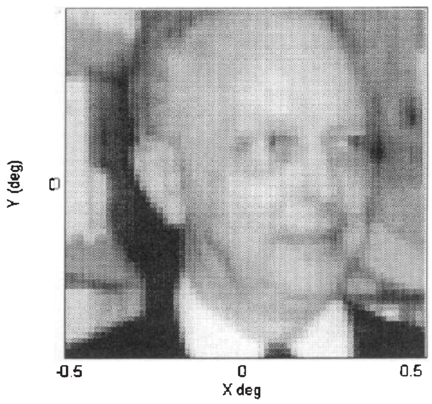
(*b*)

Figure 8. PSFs of the astigmatic observer obtained from (*a*) the Fourier characterization and (*b*) the Gabor characterization.

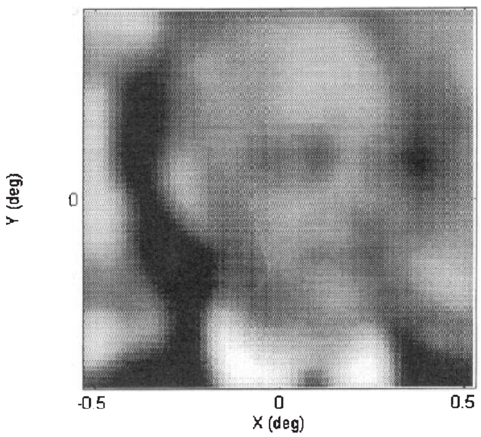
number of spatial positions considered for every frequency, increasing the size of the coefficients matrix of the GT.

As our purpose is only to show the potential of the characterization presented to perform this task, it will be sufficient to demonstrate that the GSF can include inhomogeneous behaviours. This fact will be examined with a synthesized example

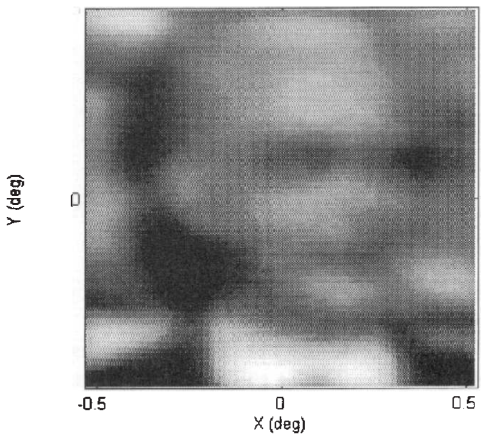
(*a*)



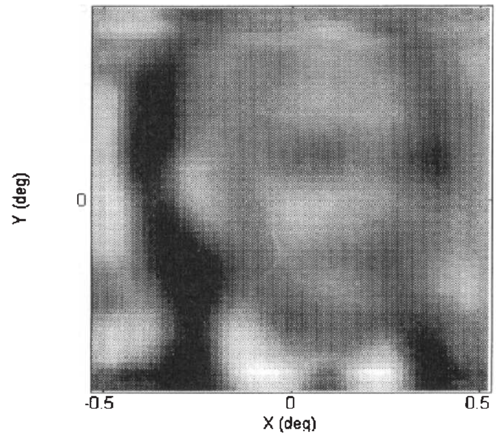
(*b*)



(*c*)



(*d*)



(*e*)

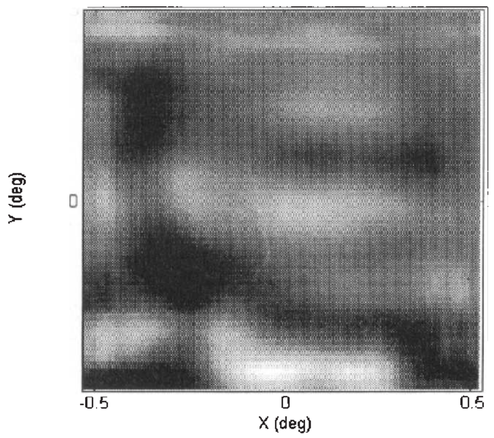


Figure 9. Natural image filtering. (*a*) Shows the original natural image (Prof. Givens 1964 photo from the Gatlinburg Conference on Numerical Algebra). (*b*) Shows the natural image weighted by the GSF of the emmetropic observer and (*c*) shows the result using the CSF. (*d*) Displays the natural image weighted by the GSF of the astigmatic observer. Note the different behaviour of both observers. The astigmatic observer blurs the vertical borders (horizontal frequencies) and relatively enhances the horizontal borders (vertical frequencies). This can be seen in both Gabor and Fourier characterizations. (*e*) shows the natural image weighted by the CSF of the astigmatic observer.

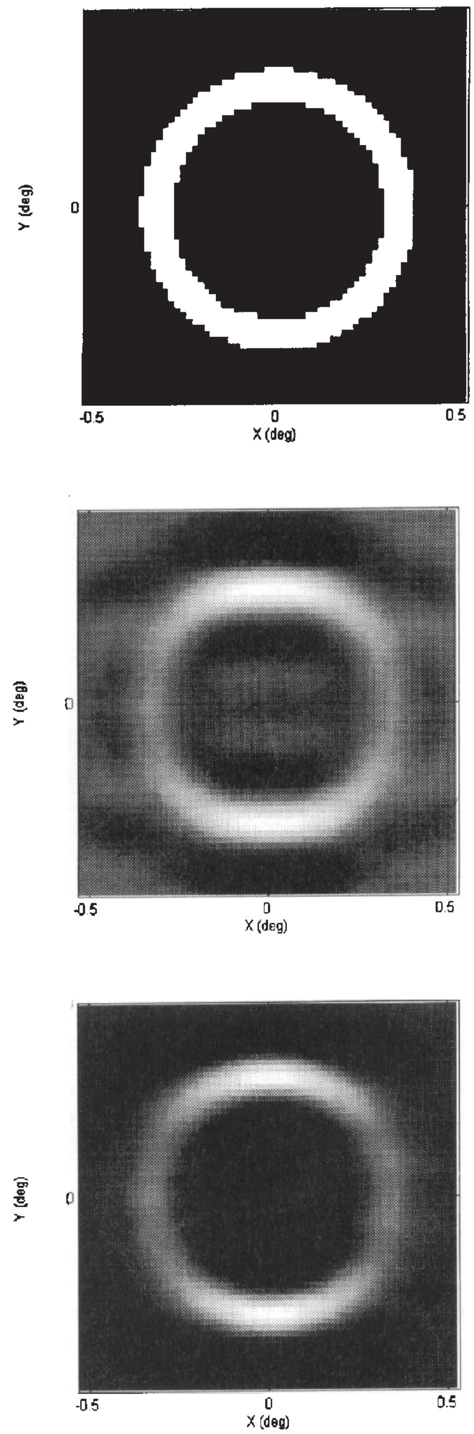
of non-homogeneity in the foveal region, which allows us to use the algorithm already developed. The introduction of spatial asymmetries means loss of uniformity in the spatial coefficients of a given submatrix of a particular frequency. In other words, it means imposing spatially varying sensitivity on every frequency component.

The weighting function of the simulated observer is generated by substituting the GSF values of the observer RZ in the spatial positions corresponding to the

(*a*)

(*b*)

(*c*)



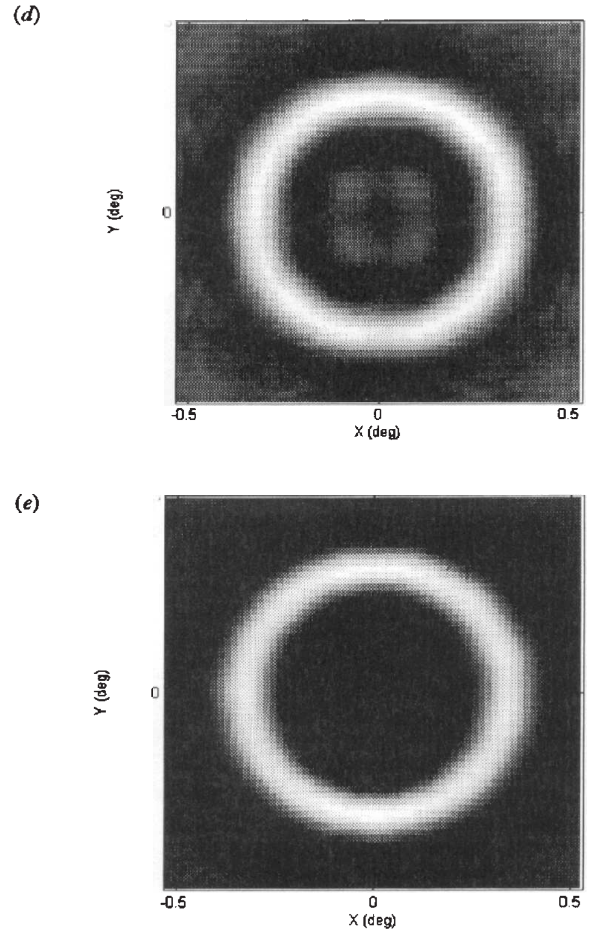


Figure 10. Circular test filtering. (*a*) Depicts the original circular test. (*b*) Shows the circular test weighted by the GSF of the astigmatic observer and (*c*) shows the analogue result using the CSF. (*d*) Displays the circular test processed by the GSF of the emmetropic observer and (*e*) shows the result in the Fourier characterization for the emmetropic observer.

upper part of the image in the GSF of the observer JM. This gives astigmatic behaviour for $y > 0$ and emmetropic behaviour for $y < 0$. The result of the synthesized weighting function is shown in figure 11, and the result of the perception simulation of the circular test in figure 12.

The different degree of astigmatism, in the different zones considered, is apparent in figure 12. The definition at the borders of the test is limited to a small angle in the upper half (strong astigmatism), while in the lower half the angular range is bigger. The transition from emmetropic to astigmatic behaviour is smooth because the differences in attenuation between spatial zones corresponding to each of the filters have been smoothed (compare figures 11 and 6). The superposition of two functions, of equal frequency, in close spatial positions and with very different

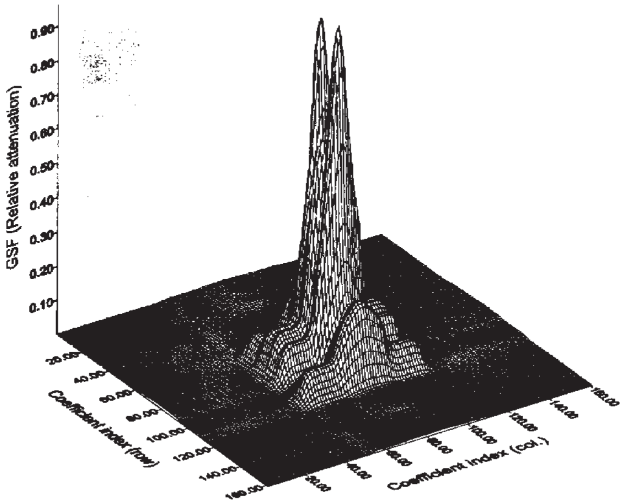


Figure 11. Synthetic GSF obtained mixing the weighting functions of figures 6 (*a*) and (*b*) in different spatial positions.

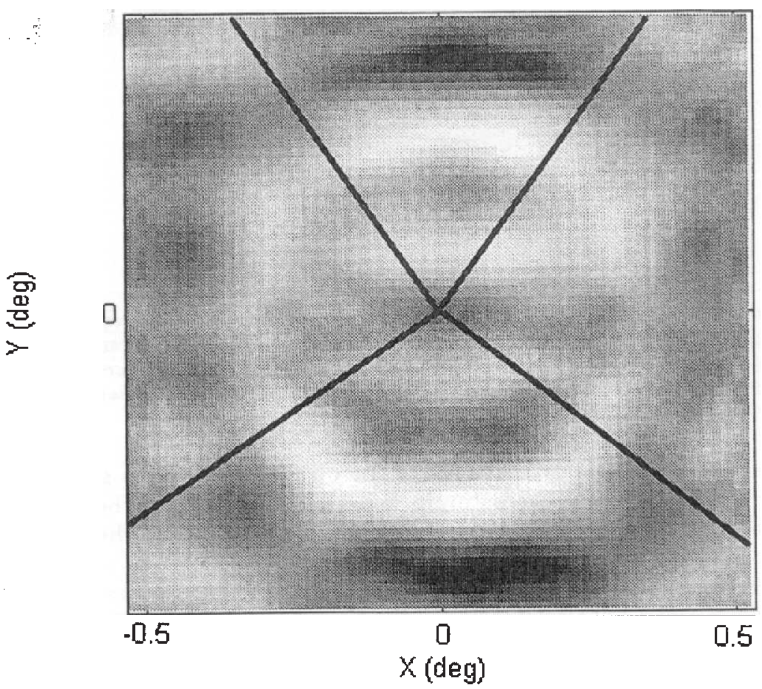


Figure 12. Circular test processed by the synthetic weighting function. Note the astigmatic behaviour in the upper half of the figure and the emmetropic behaviour in the lower half, evidenced by the difference of the angle of sharp discrimination.

attenuation, implies the existence of noise in the transition zone originated by non-cancelled oscillations due to the partial non-orthogonality of the base function of the same frequency. This fact imposes a restriction on the characterization of systems with too sharp spatial variations. This is not a problem in the visual system case where a smooth spatial variation is reported [39, 40]. In our synthetic example we must obtain a smooth variation between spatial zones with different attenuation. In this case we performed the smoothing by convolution of the synthesized filter with a Gaussian function.

This example does not attempt to be a model for the spatial variations of the visual system, but it simply manifests that such variations can be included perfectly well in a characterization such as the one presented. If the desired local Fourier weighting functions (CSF_i) are known, we can take advantage of the relation between the CSF and its analogue in the Gabor domain, the GSF (equation 4 *a*), to calculate the conjoint homogeneous weighting functions GSF_i from their Fourier equivalents CSF_i, and afterwards combine the coefficients from the different GSF_i in the desired spatial positions in order to generate the single non-homogeneous weighting function in the conjoint domain. In this way, we achieve a straight characterization of the spatially variant system: instead of applying multiple convolution kernels in different spatial positions, a single conjoint attenuation process takes into account all the non-homogeneities of the system.

To conclude, we would like to restate here that the conjoint weighting idea and the equations stated are not restricted by the selection of either the basis or the transformation. Therefore, it is also possible to work on defining a weighting function according to either another basis or another space/spatial frequency linear transformation, which may be considered to be more suitable or to be a better representation of the physiology of vision.

References

[1] Kersten, D., 1987, *J. opt. Soc. Am.* A, **4**, 2395.
[2] Gonzalez, R. C., and Wintz, P., 1987, Digital Image Processing (Reading, MA: Addison-Wesley), chap. 6.
[3] Field, D., 1987, *J. opt. Soc. Am.* A, **4**, 2379.
[4] Campbell, F. W., and Green, D., 1965; *J. Physiol.*, **181**, 576.
[5] de Valois, R. L., Albrecht, D. G., and Thorell, L. G., 1982, *Vision Res.*, **22**, 545.
[6] Webster, M. A., and de Valois, R. L., 1985, *J. opt. Soc. Am.* A, **2**, 1124.
[7] Marcelja, S., 1980, *J. opt. Soc. Am.*, **70**, 1297.
[8] Daugman, J. G., 1980, *Vision Res.*, **20**, 847.
[9] Daugman, J. G., 1983, *IEEE Trans. Syst., Man, Cybern.*, **13**, 882.
[10] Daugman, J. G., 1984, *Vision Res.*, **24**, 891.
[11] Harvey, L. O., and Doan, V. V., 1990, *J. Opt. Soc. Am.*, **7**, 116.
[12] Daugman, J. G., 1985, *J. opt. Soc. Am.* A, **2**, 1160.
[13] Daugman, J. G., 1988, *IEEE Trans. Acoust., Speech, Signal Process.*, **36**, 1169.
[14] Porat, M., and Zeevi, Y. Y., 1988, *IEEE Trans. Pattern Anal. machine Intell.*, **10**, 452.
[15] Field, D. J., Hayes, A., and Hess, R. F., 1993, *Vision Res.*, **33**, 173.
[16] Watson, A. B., 1983, *Physical and Biological Processing of Images*, edited by Braddick and Sleigh (Berlin: Springer), pp. 100–114.
[17] Watson, A. B., 1987, *Comput. Vision Graph. Image Process.*, **39**, 311.
[18] Watson, A. B., 1987, *J. Opt. Soc. Am.* A, **4**, 2401.
[19] Stork, D. G., and Wilson, H. R., 1990, *J. opt. Soc. Am.* A, **7**, 1368.
[20] Daubechies, I., 1988, *Commun. pure appl. Math.*, **41**, 909.

[21] MALLAT, S. G., 1989, *Trans. Am. math. Soc.*, **315**, 69.
[22] CHUI, C. K., 1992, *An Introduction to Wavelets* (London: Academic Press).
[23] MALLAT, S. G., 1989, *IEEE Trans. Pattern Anal. machine Intell.* **7**, 674.
[24] Special Issue on Wavelet Analysis, 1992, *IEEE Trans. Inf. Theory*, **38** (2).
[25] GABOR, D., 1946, *JIEE*, **93**, 429.
[26] NAVARRO, R., and TABERNERO, A., 1991, *Multidim. Sys. Signal Process.*, **2**, 421.
[27] BASTIAANS, M. J., 1982, *Optica Acta*, **29**, 1223.
[28] EBRAHIMI, T., and KUNT, M., 1991, *Opt. Eng.*, **30**, 873.
[29] COHEN, A., and FROMENT, J., 1992, *Wavelets and Applications*, edited by Y. Meyer (Berlin: Springer), pp. 181–206.
[30] ANTONINI, M., BARLAUD, M., and MATHIEU, P., 1992, *Wavelets and Applications*, edited by Y. Meyer (Berlin: Springer), pp. 160–174.
[31] BOVIK, A. C., CLARK, M., and GEISLER, W. S., 1990, *IEEE Trans. Pattern Anal. Machine Intell.*, **12**, 55.
[32] GAUDART, L., CREBASSA, J., and PETRAKIAN, J. P., 1993, *Appl. Optics*, **32**, 4119.
[33] LANDY, M. S., and BERGEN, J. R., 1991, *Vision Res.*, **31**, 679.
[34] POLAT, U., and SAGI, D., 1994, *Vision Res.*, **34**, 73.
[35] SALEH, B. E. A., 1982, *Applications of Optical Fourier Transforms*, edited by H. Stark (New York: Academic Press), chap. 10.
[36] PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A., and VETTERLING, W. T., 1992, *Numerical Recipes in C: the Art of Scientific Computing* (Cambridge: Cambridge University Press), chap. 2.
[37] MALO, J., FELIPE, A., LUQUE, M. J., and ARTIGAS, J. M., 1994, *J. Optics*, **25**, 93.
[38] PROAKIS, J. G., and MANOLAKIS, D. G., 1992, *Digital Signal Processing. Principles, Algorithms and Applications* (New York: Macmillan), chap. 6, p. 450.
[39] KELLY, D. H., 1985, *Vision Res.*, **25**, 1895.
[40] NAVARRO, R., ARTAL, P., and WILLIAMS, D. R., 1993, *J. opt. Soc. Am.* A, **10**, 201.

A.3 The role of the perceptual contrast non-linearities in image transform quantization

Image and Vision Computing. (Aceptado Abril 1999)

The role of perceptual contrast non-linearities in image transform quantization [1]

J. Malo [a], F. Ferri [b], J. Albert [b], J. Soret [b] and J.M. Artigas [a]

[a] *Dpt. d'Òptica*

[b] *Dpt. d'Informàtica i Electrònica*

Universitat de València

C/ Dr. Moliner 50. 46100. Burjassot, València, Spain

e-mail: Jesus.Malo@uv.es

The conventional quantizer design based on *average error* minimization over a training set does not guarantee a good subjective behaviour on individual images even if perceptual metrics are used.

In this work a novel criterion for transform coder design is analyzed in depth. Its aim is to bound the perceptual distortion in each individual quantization according to a non-linear model of early human vision. A common comparison framework is presented to describe the qualitative behaviour of the optimal quantizers under the proposed criterion and the conventional rate-distortion based criterion. Several underlying metrics, with and without perceptual non-linearities, are used with both criteria. Analytical results show that the proposed design criterion gives rise to a JPEG-like quantization if a simple linear metric is used. Experimental results show that significant improvements over the perceptually weighted rate-distortion approach are obtained if a more meaningful non-linear metric is used.

Key words: Image Coding. Perceptual Quantization. Non-linear Perception Model.

1 Introduction

According to the well-known image transform coding paradigm, an appropriate transformation of the signal is carried out prior to the quantization stage to simplify the quantizer design. Given a certain transform (usually a local frequency transform), two intimately related problems must be solved. The first problem is to decide how to distribute quantization levels in the amplitude

[1] Partially supported by CICYT projects TIC98-677-C02-02 and TIC 1FD97-279.

range of each transform coefficient. The second problem is to determine the number of quantization levels that will be devoted to each coefficient. The first is usually known as *1D quantizer design* and the second is often referred to as *bit allocation.*

It is widely accepted that for image coding applications that are judged by a human observer, the properties and limitations of the human visual system (HVS) have to be taken into account in order to achieve better subjective results [1]. In transform-based coding, the sensitivity of the human viewer to the transform basis functions should be taken into account to adapt the quantization step size for each coefficient to minimize the perceived distortion [2, 3]. As stated by Watson [**?**], there are two ways to exploit the HVS characteristics in image coding: *a)* the usual *signal-based* way, which consists of adapting the conventional approach based on minimizing the *average error* [4] through the appropriate perceptual distortion metric, and *b)* an alternative *fully-perceptual* way, which mimics the way the HVS encodes the visual information.

The HVS characteristics been applied to different approaches in a number of ways. While the HVS sensitivity frequency dependency has been widely used for bit allocation [4–14], the corresponding amplitude dependencies have rarely been used, and only under the fully-perceptual paradigm [15–17].

Signal-based quantizers that minimize the mean square error have some drawbacks for individual images [18] even if the error measure is perceptually weighted [10]. This is mainly because the statistics of the training set may bias the encoding effort towards the most populated regions in order to minimize the average distortion. Under these conditions some values of amplitude for certain coefficients of importance for the HVS may become badly represented. This effect is particularly important at high compression ratios when the number of reproduction levels is dramatically reduced and the probability of badly represented amplitudes increases. This suggests that the use of an *average error* criterion may not be the best choice given a particular HVS model.

The objective of this work is to assess the relative merits of a perceptually weighted signal-based coding scheme when a HVS model including frequency and amplitude dependencies is considered. To show that this is not the best *subjective* way of solving the problem, an alternative to the weighted MSE criterion is proposed: the *maximum perceptual error* (MPE) criterion. While the MSE implies a minimization over a training set, the MPE consists of setting limits on the amount of distortion permitted for every frequency and amplitude taking into account the HVS.

The main goal is to compare these two design criteria when using different HVS models. Moreover, the particular effects of introducing the amplitude non-linearities is analyzed in depth in both cases.

The paper is organized as follows. The selected perceptual metric including frequency and amplitude properties of the HVS is introduced in Section 2. In section 3, this metric is used to obtain particular quantizers from the asymptotic results of the rate-distortion theory, based on the MSE criterion. The alternative MPE criterion is presented in section 4. In section 5, the performance of the optimal quantizers using both design criteria with the same perceptual metrics is compared. To conclude, some final remarks are presented in section 6.

2 Frequency and amplitude-dependent perceptual metric

According to currently accepted perception models [19–22], the HVS maps the input spatial patterns, $\mathbf{A}$, onto a frequency feature space through a set of m local frequency filters, $\mathbf{a} = T[\mathbf{A}]$, with components a_f, where $f = 1, \ldots, m$. Then, a non-linear log-like transform, R, is applied to yield the so-called response representation, $\mathbf{r} = R[\mathbf{a}]$, with components r_f.

The perceptual metric in the feature space, $W(\mathbf{a})$, can be obtained from an appropriate contrast response model. The metric is related to the non-linearities of the response R and, under certain assumptions, can be directly computed from its gradient as $W(\mathbf{a}) = \nabla R(\mathbf{a})^T \cdot \nabla R(\mathbf{a})$ [23]. It can also be empirically obtained from the HVS discrimination abilities. In the feature space, two patterns are perceived as different if the distance between them is above a certain discrimination threshold value, τ. The vector difference between two just distinguishable patterns, $\Delta\mathbf{a}^\star(\mathbf{a})$, is called *just noticeable difference* (JND), and the square perceptual distance, D^2, under the *ideal observer assumption* [21, 22] can be written as:

$$D(\mathbf{a}, \mathbf{a} + \Delta\mathbf{a}^\star(\mathbf{a}))^2 \;=\; \Delta\mathbf{a}^\star(\mathbf{a})^T \cdot W(\mathbf{a}) \cdot \Delta\mathbf{a}^\star(\mathbf{a}) \;=\; \tau^2 \qquad (1)$$

All the elements of the metric can be obtained with an appropriate number of JND data at a point, $\mathbf{a}$. If the interaction among the filter outputs is neglected, W will be diagonal. In this case, each non-zero element, W_f, will not depend on $a_{f'}$ for $f' \neq f$ [21, 23]. From eq. 1, each diagonal element will be inversely proportional to the square of the JND in the corresponding axis. Using a fit of the amplitude JND data, a previously tested metric [23, 24] is obtained:

$$W_f(a_f) = \tau^2 \Delta a_f^\star(a_f)^{-2} = \tau^2 \left(CSF_f^{\,-1} + \frac{\frac{a_f}{L}\left(k_f\left(\frac{a_f}{L}\right)^{n_f} - CSF_f^{\,-1}\right)}{(k_f CSF_f)^{-1/n_f} + \frac{a_f}{L}} \right)^{-2} \quad (2$$

where L is the local mean luminance, CSF_f is the *Contrast Sensitivity Function*, which can be computed with the expressions given by Kelly [25], Nill [26] or Nygan [12, 13], and k_f and n_f are the empirical dependency on frequency of the JND parameters [24].

It is worth pointing out that eq. 2 includes the effect of non-linearities as a correction of the linear threshold behavior represented by the CSF band-pass filter. If the threshold behavior, $a_f \to 0$, is assumed to be valid for suprathreshold amplitudes (as is done in simple linear models) the amplitude dependent term in eq. 2 vanishes and the CSF-based metric [26, 27] is obtained.

If a given coding scheme uses some particular transform basis, specific incremental threshold data should be used [2, 3]. However, as the impulse responses of the perceptual filter bank, T, closely resemble the basis functions used in image coding transforms [28, 29] a similar behavior can be expected [2]. Therefore the metric W can be considered a reasonable approximation with any generic local frequency transform. In some cases, it is also possible to obtain the behavior in the desired domain from the functions defined in other domains [26, 31].

3 Coder design through perceptual MSE minimization

The standard approach to transform coder design is rooted at the minimization of the mean square error, $\overline{D}_f^2$, between the original amplitude of each transform coefficient, a_f, and its quantized version. Using a generic perceptual metric, W, to measure the elemental distortions, it holds that:

$$\overline{D}_f^2 = \sum_{i=1}^{N_f} \int_{R_i} (a_f - {a_f}^i)^2 \, W_f({a_f}^i) \, p(a_f) \, da_f \qquad (3)$$

where N_f is the number of quantization levels for that coefficient, ${a_f}^i$ is the i-th quantization level, R_i is the quantization region corresponding to the i-th level and $p(a_f)$ is the probability density function (*pdf*) of a_f.

Using the asymptotic results of Gish et al. [32] and Yamada et al. [33] for general non-euclidean metrics, the Bennett distortion integral [34] can be obtained in this case:

$$\overline{D}_f^2 = \frac{1}{12 N_f^2} \int \frac{W_f(a_f) \, p(a_f)}{\lambda_f(a_f)^2} \, da_f \qquad (4)$$

[2] For instance, the same kind of amplitude non-linearity or absolute detection threshold for sinusoids and DCT basis functions is usually assumed [2, 5, 16, 30].

This expression relates the average distortion of a coefficient using the proposed metric with the density of quantization levels for that coefficient, $\lambda_f(a_f)$. The optimal 1D quantizer with the proposed metric can be obtained using the Hölder inequality in the standard way [4]:

$$\lambda_{f\,opt}(a_f) = \frac{(W_f(a_f)\,p(a_f))^{1/3}}{\int (W_f(a_f)\,p(a_f))^{1/3}\,da_f} \tag{5}$$

This result only differs in the factor $W_f(a_f)$ from the well-known euclidean result which is proportional to the cube root of the *pdf* [4, 35, 36].

The inclusion of the amplitude dependence of the metric also has significant effects on the bit allocation results. As the global squared distortion is given by the sum of the squared distortions coming from each coefficient, the individual distortions should be kept below a certain threshold in order to keep the global distortion as low as possible. This requirement determines a non-uniform bit allocation in the transform domain because more bits should be used to encode the more demanding coefficients [4, 37]. The intrinsic demand of each coefficient is given by its contribution to the total distortion. Substituting $\lambda_{f\,opt}$ in equation 4, the distortion for each optimal quantizer is:

$$\overline{D}^2_{f\,opt} = \frac{\sigma_f^2}{12N_f^2}\left(\int (W_f(\sigma_f a_f)\,\tilde{p}(a_f))^{1/3}\,da_f\right)^3 = \frac{\sigma_f^2}{N_f^2}\cdot H_f \tag{6}$$

where $\tilde{p}(a_f)$ is the normalized unit-variance *pdf* of the coefficient; the variance, σ_f^2, and the metric dependent parameter, H_f, can be seen as the weighting factors that control the intrinsic relevance of each coefficient. The optimal bit allocation is obtained imposing the same average distortion for each f [4]. In our case, from the optimal distortion results of equation 6, we obtain:

$$b_{f\,opt} = \frac{B}{m} + \frac{1}{2}\log_2(\sigma_f^2\cdot H_f) - \frac{1}{2m}\sum_{f=1}^{m}\log_2(\sigma_f^2\cdot H_f) \tag{7}$$

where $b_f = \log_2 N_f$ is the number of bits used to encode the coefficient f; and B is the total number of available bits.

The final effect of the amplitude non-linearities is to weight the *pdf* in the 1D quantizer design and to introduce a new factor in the bit allocation expressions. These expressions generalize the previously reported results in the context of image transform coding [4, 9, 10] to include amplitude dependencies of the perceptual metric.

If only the frequency dependency of HVS sensitivity is taken into account (as is done in simple linear models), the metric $W_f(a_f)$ will just be a function of frequency related to the CSF. Thus, W_f can be taken from the integrals in eqs. 5 and 6 to give an explicit perceptual frequency weight term in the bit allocation expression [4, 9, 10]. In this case, the 1D quantizers will only depend on the signal statistics, and the density of quantization levels is then proportional to the cube root of the *pdf* [4, 35, 36]. Band-pass weights have been heuristically used in several coding schemes [6, 11, 12] to introduce simple HVS features in the euclidean (purely statistical) design.

Although all these results only hold in the high resolution case, these analytical expressions clarify the effect of the the metric and often turn out to be a reasonable approximation even in the medium to low resolution cases.

4 Coder design through Maximum Perceptual Error bounding

In the different MSE-based approaches, the final quantizer may show a perceptually uneven distribution of the quantization levels depending on the statistics of the training set. The accumulation of quantization levels in the more probable regions which is performed to minimize the average perceptual error does not ensure good behavior on a particular image.

In order to prevent high perceptual errors on individual images coming from outlier coefficient values, the coder should be designed to bound the maximum perceptual error (MPE) for every frequency and amplitude. While the MSE criterion minimizes the *average* distortion, $\overline{D}_f^2$, the proposed MPE criterion bounds the *maximum individual* distortion, $\widehat{D}_f^2(a_f)$.

If a given coefficient is represented by N_f quantization levels distributed according to a density, $\lambda_f(a_f)$, the maximum euclidean quantization error at an amplitude, a_f, will be bounded by half the euclidean distance between two levels:

$$\Delta a_f(a_f) \leq \frac{1}{2N_f\,\lambda_f(a_f)} \tag{8}$$

Assuming a generic diagonal frequency and amplitude-dependent metric, the MPE at that amplitude will be related to the metric and the density of levels:

$$\widehat{D}_f^2(a_f) = \frac{1}{2N_f\,\lambda_f(a_f)} \cdot W_f(a_f) \cdot \frac{1}{2N_f\,\lambda_f(a_f)} = \frac{W_f(a_f)}{4N_f^2\,\lambda_f^2(a_f)} \tag{9}$$

According to this, the maximum perceptual distortion bound is constant over the amplitude range only if the point density varies as the square root of the metric,

$$\lambda_{f\,opt}(a_f) \;=\; \frac{W_f(a_f)^{1/2}}{\int W_f(a_f)^{1/2} da_f} \tag{10}$$

With these optimal densities, the maximum perceptual distortion in each coefficient will depend on the number of allocated levels and on the integrated value of the metric:

$$\widehat{D}^2_{f\,opt} \;=\; \frac{1}{4N_f^2}\left(\int W_f(a_f)^{1/2} da_f\right)^2 \tag{11}$$

Fixing the same maximum distortion for each coefficient, $\widehat{D}^2_{f opt} \;=\; k^2$, and solving for N_f, the optimal bit allocation is obtained:

$$b_{f\,opt} = \log_2 N_f = \log_2\left(\frac{1}{2k}\int W_f(a_f)^{1/2}\, da_f\right) \tag{12}$$

The expressions 10-12 constitute an alternative to the expressions 5-7 that are optimal in an average sense. The MPE criterion implies a perceptually uniform distribution of the available quantization levels in the feature space, and this only depends on the perceptual metric and not on the signal statistics.

A practical side effect of the fully-perceptual MPE criterion is that it does not need an off-line training stage. By its nature, it should be well-behaved over a wide range of natural images. The independence of the statistics of the images to be encoded should not be considered as rigidity: if the underlying perception model is adaptive, as are non-linear models with regard to linear models, the final quantization will be input-dependent.

It is interesting to point out that the optimal 1D quantizers in the MSE sense, eq. 5, reduce to the MPE optimal quantizer of eq. 10, if $p(a_f) \propto W_f(a_f)^{1/2}$. The differences in the placement of quantization levels will depend on the shape of the *pdf*s with regard to the shape of the metric. *Pdf*s which are too sharp may bias the distribution of quantization levels giving rise to a poor representation of some potentially important regions from a perceptual point of view. As a consequence, some subjectively relevant details of a particular image may be lost with the MSE approach even if the appropriate perceptual metric is taken into account.

The presented MPE formulation is qualitatively related to other fully perceptual approaches that have also taken the perceptual amplitude non-linearities of the HVS into account [15–17, 38]. In particular, the qualitative quantizer design of refs. [17, 38] is ruled by a sensitivity function, the *Information Allocation Function* (IAF), which is inversely proportional to the JNDs as the square root of the perceptual metric of section 2.

The equivalence between these fully perceptual approaches can be seen in their common background perception model, in the similarity of their underlying functions [3], and in their similar applications: these models have been applied to define perceptually meaningful distortion metrics [23, 24, 30, 40]. The IAF model has also been successfully used to perceptually match other signal processing algorithms, such as adaptive motion estimation [38, 41].

A particularly interesting quantizer is obtained after neglecting the amplitude-dependent term in the metric (eq. 2). In this case, uniform quantizers are obtained for each coefficient. The number of quantization levels per coefficient is simply proportional to the CSF, which is one of the recommended options included in the JPEG standard [5, 7, 8, 14].

5 Results and discussion

Experiments to compare the results of the different schemes for natural images at different compression ratios were carried out. A total of five quantizers were compared to explore the effect of perceptual amplitude non-linearities under the different design criteria:

- MSE-E: mean square error criterion with a euclidean metric, the standard *Mean Square Error* approach [4].
- MSE-F: mean square error criterion with a frequency-dependent perceptual metric [4, 6, 9–12].
- MSE-FA: mean square error criterion with a frequency and amplitude-depende perceptual metric.
- MPE-F: maximum perceptual error criterion with a frequency-dependent perceptual metric. (the CSF-based JPEG-like scheme) [5, 7, 8, 14]
- MPE-FA: maximum perceptual error criterion with a frequency and amplitude dependent perceptual metric.

In the MSE approaches, the perceptual features are gradually introduced to the average squared distortion measure combined with the statistical properties of

[3] Watson [16] uses the Legge fit of the incremental threshold data of gratings [39]. The noise-adaptive CSF used in ref. [15] is obtained from incremental thresholds of gratings over white noise and is quite similar to the IAF obtained from contrast incremental thresholds of gratings (see Fig.4.a of [15] and Fig.1 of [24]).

the signal. In the MPE approaches, the difference consists of whether or not the amplitude non-linearities are considered.

5.1 Implementation details

The training set included 60, 256×256, 8-bit monochrome video conference-like images (head-and-shoulders). The quantization block size was set to 16×16, so 15360 sample blocks were considered to estimate the amplitude *pdfs* of the MSE approach. Figure 1 shows the relevant statistical factors in the MSE approach. Figure 1.a shows the *pdf*s for some low, medium and high frequency coefficients (4, 8 and 16 cycl/deg). Figure 1.b shows the variance of each coefficient, σ_f^2. The factor H_f in the MSE-E case, eq. 6, depends only on the shape of the normalized unit-variance *pdf*s and is almost frequency-independent as reported in the literature [4].

The analytical results of sections 3 and 4 are strictly applicable under the high rate approximation [4, 34] to ensure that the *pdf*s and the metric are smooth enough to be considered constant in each quantization step. To design the actual low resolution MSE quantizers for the experiments, the LBG method [35, 42] was used with the appropriate metric. In each case, the iterative method was initialized with the asymptotic results. The final quantizers were quite consistent with the asymptotic assumptions. In the MPE-FA case, the 1D quantizer for each frequency was obtained from N_f uniformly distributed quantization levels mapped to the amplitude domain through the inverse of the proper companding function [4, 17, 38]. In the JPEG-like case, the quantization step size was set constant.

The actual bit allocation in all schemes was obtained through a greedy integer-constrained algorithm [4] based on the sequential allocation of one level to the coefficient with the largest distortion in each iteration. In the MSE and MPE cases, the distortion is given by the eqs. 6 and 11, respectively. The DC coefficient is always separately encoded using DPCM and will not be considered throughout the discussion sections.

5.2 Quantizer results

Figure 2 shows the bit allocation results for the compared approaches. In the MSE-E approach, the bit allocation is directly given by the variance. In the MSE-F approach, the CSF-based frequency weight cuts the high frequency tail of the variance curve. The MSE-FA reduces this restriction to a certain extent because the consideration of the amplitude non-linearities widens the band of perceptual interest. The same effect can also be seen in the fully perceptual approaches: while the MPE-F bit allocation is more concentrated in the low frequency region, the MPE-FA does include the high frequency region. However, this does not mean that any high frequency contribution is going to be preserved in the MPE-FA scheme. It will depend on its amplitude and on the shape of the density of quantization levels for that frequency.

The density of quantization levels of the different approaches for different coefficients (4, 8 and 16 cycl/deg) is shown in Figure 3. As expected from the statistical results of Figure 1, in the MSE-E and MSE-F cases, the quantization levels are more and more concentrated around zero as the frequency increases. In the MPE-FA case, the density is sharper than the cubic root of the *pdf* for low frequencies. However, the square root of the metric becomes flat as the frequency increases, i.e., the perceptual relevance of high amplitude regions increases with frequency. In the MSE-FA case, the effect of the amplitude perceptual weighting is larger in the low frequencies when $W_f(a_f)^{1/2}$ is sharper (Figure 3.a). As the metric becomes flat, the main amplitude dependency in eq. 5 is due to the *pdf* and the MSE-FA density tends to the MSE-E density.

5.3 Comparison strategy: the density surface in frequency and amplitude

The final effect of the different design criteria, MSE and MPE, is a different distribution of the total number of available reproduction levels in the encoding domain. To visualize and compare such an effect for the different approaches, the definition of a *density surface* in the frequency and amplitude plane, $\Lambda_f(a_f)$, is proposed. This surface is obtained by scaling the density of quantization levels of each coefficient by the allocated levels in that coefficient:

$$\Lambda_f(a_f) = N_f \cdot \lambda_f(a_f) \tag{13}$$

It can be interpreted as a density of quantization levels in the frequency and amplitude plane. This is simply a generalization of the role played by the the perceptual metric in the MPE approaches, for which the density surface is $\Lambda_f(a_f) \propto W_f(a_f)^{1/2}$. The density surface has the same properties as the square root of the metric in the MPE approaches:

- The location of the quantization levels of one coefficient are given by the cut of the surface for that coefficient (as in eq. 10):

$$\lambda_f(a_f) = \frac{\Lambda_f(a_f)}{\int \Lambda_f(a_f)\, da_f} \tag{14}$$

- The bit allocation is given by the integral of the density surface in amplitude (as in eq. 12):

$$b_f = \log_2 \left(\int \Lambda_f(a_f)\, da_f \right) \tag{15}$$

In this way, the relative importance given to the different regions of the encoding domain by the different design approaches is described by a single common function, $\Lambda_f(a_f)$, which completely determines the transform coder design (λ_f and b_f). The different design approaches can be understood as different ways of defining the density surface.

Figure 4 shows the density surface, $\Lambda_f(a_f)$, in MSE-E, MSE-F, MSE-FA and MPE-FA cases, along with contour views of the same functions. The integrals of the density surfaces in amplitude give the number of levels per coefficient for each approach (Figure 2).

These figures highlight the joint effect of the signal statistics and the perceptual metric in the quantizer design. In the first case, MSE-E, the density surface is fully determined by the signal statistics. The second and the third cases show the effect of the progressive introduction of frequency and amplitude perceptual features, MSE-F and MSE-FA. Figure 4.d shows the amplitude-independent surface that accounts for the simple CSF model extended for suprathreshold amplitudes. It represents the uneven bit allocation and uniform quantization of a JPEG-like approach. The last figure shows the fully perceptual MPE-FA surface: the square root of the metric. This function is inversely proportional to the contrast incremental thresholds for each frequency and amplitude (eq. 2), so it reduces to the IAF of refs. $[17, 24, 38, 41]$. In the MPE cases, the density surface is independent of the signal statistics and isolates the perceptual factors that have been progressively added in the MSE-F and MSE-FA cases.

In the MSE-F case, the frequency dependency of $W_f^{1/2}$ for low amplitudes (the CSF) cuts off the high frequency tail of the MSE-E density surface. In the MSE-FA case, the contrast non-linearities restrict the encoding effort to lower amplitudes for lower frequencies. The encoding effort is slightly extended to higher amplitudes for high frequencies following the trends of the purely perceptual function.

The differences between the MSE-FA and the MPE-FA approaches depend exclusively on the design criterion. In spite of the qualitative similarity of the density surfaces due to the introduction of the perceptual metric in the MSE approach, the actual distribution of reproduction levels is still very different. From the perceptual point of view, the encoding effort in the MSE-FA approach is still too focused on the low amplitude range, especially in high frequencies. In this way, contributions of relatively high contrast in the high frequency region will be badly represented giving rise to annoying artifacts.

5.4 Decoded images

Examples of the compression results on the standard images of Lena and Barbara are shown in Figures 5 and 6. The MSE schemes and the MPE schemes were used at the same compression ratio (0.5 bpp). These images were not included in the corresponding training process.

The experimental results show three basic trends:

- In the MSE criterion, the use of a perceptual frequency and amplitude-dependent metric, MSE-FA, does not significantly improve the final results on individual images as compared with frequency-only perceptual metrics, MSE-F, or even with euclidean metric, MSE-E.
- In the MPE criterion, the effect of the amplitude non-linearities is not negligible. As previously reported [17], the CSF-based JPEG results (MPE-F) are significantly improved if these properties are taken into account.
- The MPE-FA approach gives sharper edges and better overall subjective results, especially in the high frequency details.

These results suggest that the problems of the MSE approaches are due to the improper distribution of the quantization levels in the high frequency coefficients: they are so concentrated around zero that the less probable high amplitudes are always badly represented. This implies that the high amplitude contributions in high frequency coefficients in individual images such as Barbara's clothes and Lena's hair or eye details are severely quantized. The problem with CSF-based JPEG is its band restriction in the high frequency region in such a way that these coefficients are discarded at high compression ratios. Compare the region between 8 and 16 cycl/deg in Figures 4.f to 4.j.

The MPE-FA scheme allocates certain bits in the high frequency region but spreads the quantization levels over a relatively extensive amplitude region. In this way, the first quantization level is high enough to discard irrelevant contributions, but the levels positioned at higher amplitudes ensure an approximate representation of high contrast and high frequency contributions.

These results clarify the relative importance of the different factors affecting the quantizer design. On one hand, the relevance of the distortion metric depends on the design criterion: while the MSE approach, ruled by the signal statistics, does not exploit the eventual advantages of more sophisticated metrics, the results of the MPE approach significantly improve with the introduction of additional perceptual features in the metric. On the other hand, given an appropriate metric including frequency and amplitude features of contrast discrimination, the design criterion makes the difference: the important thing is not to minimize the average of a perceptually matched distortion, but either to ensure that, in a particular image, the MPE will be below a certain value. This bound completely depends on the perceptual geometry of the encoding domain and not on the signal statistics.

6 Final remarks

An alternative design criterion for transform coder design has been compared with the minimization of the average error, using distortion metrics that include frequency and amplitude perceptual features to different extents. The goal of the proposed criterion is to bound the maximum perceptual error in each coefficient regardless of the amplitude of the input. If a non-linear perception model and metric is assumed, this MPE criterion involves a non-uniform, input-dependent, quantization of the amplitude axis of the image transform.

Using a non-linear perceptual metric, the proposed MPE criterion improves the results of the MSE quantizers on natural imagery even if the MSE is provided with the same distortion metric. While the improvements of the proposed scheme with regard to the JPEG-like quantizer are due to the use of a more accurate perception model, its improvements with regard to the perceptually weighted MSE approach are due to the use of a perceptually meaningful design criterion. These results suggest that bounding the perceptual distortion in each particular block of the image may be more important than minimizing the *average* perceptual distortion over a set of images.

A practical side effect of the fully-perceptual MPE criterion is that it does not need off-line training stage. By its nature, it should be well-behaved over a wide range of natural images.

References

[1] N. Jayant, J. Johonston, and R. Safranek, "Signal compression based on models of human perception," *Proceedings IEEE*, vol. 81, no. 10, pp. 1385–1422, 1993.

[2] J. Solomon, A. Watson, and A. Ahumada, "Visibility of DCT basis functions: effects of contrast masking," in *Proceedings of Data Compression Conference, Snowbird, Utah*, IEEE Computer Society Press, pp. 361–370, 1994.

[3] A. Watson, G. Yang, J. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Transactions on Image Processing*, vol. 6, pp. 1164–1175, 1997.

[4] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Press, 1992.

[5] A. Ahumada and H. Peterson, "Luminance-model-based DCT quantization for color image compression," vol. 1666 of *Proceedings of the SPIE*, pp. 365–374, 1992.

[6] C. Hwang, S. Venkatraman, and K. Rao, "Human visual system weighted progressive image transmission using lapped orthogonal transform classified vector quantization," *Optical Engineering*, vol. 32, no. 7, pp. 1525–1530, 1993.

[7] D. LeGall, "MPEG: A video compression standard for multimedia applications," *Communications of the ACM*, vol. 34, no. 4, pp. 47–58, 1991.

[8] A. Leger, T. Omachi, and G. Wallace, "JPEG still picture compression algorithm," *Optical Engineering*, vol. 30, no. 7, pp. 947–954, 1991.

[9] B. Macq and H. Shi, "Perceptually weighted vector quantization in the DCT domain," *Electronics Letters*, vol. 29, no. 15, pp. 1382–1384, 1993.

[10] B. Macq, "Weighted optimum bit allocations to orthogonal transforms for picture coding," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 875–883, 1992.

[11] D. McLaren and D. Nguyen, "Removal of subjective redundancy from DCT-coded images," *Proceedings IEE-I*, vol. 138, no. 5, pp. 345–350, 1991.

[12] K. Nygan, K. Leong, and H. Singh, "Adaptive cosine transform coding of images in the perceptual domain," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 37, no. 11, pp. 1743–1750, 1989.

[13] K. Nygan, H. Koh, and W. Wong, "Hybrid image coding scheme incorporating human visual system characteristics," *Optical Engineering*, vol. 30, no. 7, pp. 940–946, 1991.

[14] G. Wallace, "The JPEG still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 31–43, 1991.

[15] S. Daly, "Application of a noise-adaptive Contrast Sensitivity Function to image data compression," *Optical Engineering*, vol. 29, no. 8, pp. 977–987, 1990.

[16] A. Watson, "DCT quantization matrices visually optimized for individual images," in *Human Vision, Visual Processing and Digital Display IV* (B. Rogowitz, ed.), vol. 1913, 1993.

[17] J. Malo, A. Pons, and J. Artigas, "Bit allocation algorithm for codebook design in vector quantization fully based on human visual system non-linearities for suprathreshold contrasts," *Electronics Letters*, vol. 31, no. 15, pp. 1229–1231, 1995.

[18] G. Schuster and A. Katsaggelos, *Rate-Distortion Based Video Compression*. Boston: Kluwer Academic Publishers, 1997.

[19] J. Lubin, "*The Use of Psychophysical Data and Models in the Analysis of Display System Performance*," in *Digital Images and Human Vision* (A. Watson, ed.), (Massachusetts), pp. 163–178, MIT Press, 1993.

[20] A. Watson, "Efficiency of a model human image code," *Journal of Optical Society of America A*, vol. 4, no. 12, pp. 2401–2417, 1987.

[21] A. Watson and J. Solomon, "A model of visual contrast gain control and pattern masking," *Journal of the Optical Society of America A*, vol. 14, pp. 2379–2391, 1997.

[22] H. Wilson, "Pattern discrimination, visual filters and spatial sampling irregularities," in *Computational Models of Visual Processing* (M. Landy and J. Movshon, eds.), (Massachussets), pp. 153–168, MIT Press, 1991.

[23] A. Pons, J. Malo, J. Artigas, and P. Capilla, "Image quality metric based on multidimensional contrast perception models," *Displays*, Accepted Feb. 1999.

[24] J. Malo, A. Pons, and J. Artigas, "Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain," *Image and Vision Computing*, vol. 15, no. 7, pp. 535–548, 1997.

[25] D. Kelly, "Receptive field like functions inferred from large area psychophysical measurements," *Vision Research*, vol. 25, no. 12, pp. 1895–1900, 1985.

[26] N. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Transactions on Communications*, vol. 33, pp. 551–557, 1985.

[27] L. Saghri, P. Cheatheam, and A. Habibi, "Image quality measure based on a human visual system model," *Optical Engineering*, vol. 28, no. 7, pp. 813–819, 1989.

[28] J. Daugman, "Uncertainity relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America*, vol. 2, no. 7, pp. 1160–1169, 1985.

[29] S. Marcelja, "Mathematical description of the response of simple cortical cells," *Journal of the Optical Society of America*, vol. 70, no. 11, pp. 1297–1300, 1980.

[30] A. Ahumada, "Computational image quality metrics: A review," in *Intl. Symp. Dig. of Tech. Papers, Sta. Ana CA* (J. Morreale, ed.), vol. 25 of *Proceedings of the SID*, pp. 305–308, 1993.

[31] J. Malo, A. Pons, A. Felipe, and J. Artigas, "Characterization of human visual system threshold performance by a weighting function in the Gabor domain," *Journal of Modern Optics*, vol. 44, no. 1, pp. 127–148, 1997.

[32] H. Gish and J. Pierce, "Asymptotically efficient quantizing," *IEEE Transactions on Information Theory*, vol. 14, pp. 676–683, 1968.

[33] Y. Yamada, S. Tazaki, and R. Gray, "Asymptotic performance of block quantizers with difference distortion measures," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 6–14, 1980.

[34] W. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, 1948.

[35] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 127–135, 1982.

[36] A. Gersho, "Asymptotically optimal block quantization," *IEEE Transactions on Information Theory*, vol. 25, no. 4, pp. 373–380, 1979.

[37] J. Huang and P. Schultheiss, "Block quantization of correlated gaussian random variables," *IEEE Transactions on Communications Systems*, vol. 11, no. 3, pp. 289–296, 1963.

[38] J. Malo, F. Ferri, J. Albert, and J. Artigas, "Adaptive motion estimation and video vector quantization based on spatio-temporal non-linearities of human perception," *Lecture Notes on Computer Science, Springer Verlag*, vol. 1310, pp. 454–461, 1997.

[39] G. Legge, "A power law for contrast discrimination," *Vision Research*, vol. 18, pp. 68–91, 1981.

[40] S. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision* (A. Watson, ed.), (Massachusetts), pp. 179–206, MIT Press, 1993.

[41] J. Malo, F. Ferri, J. Albert, and J. Artigas, "Splitting criterion for hierarchical motion estimation based on perceptual coding," *Electronics Letters*, vol. 34, no. 6, pp. 541–543, 1998.

[42] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, 1980.

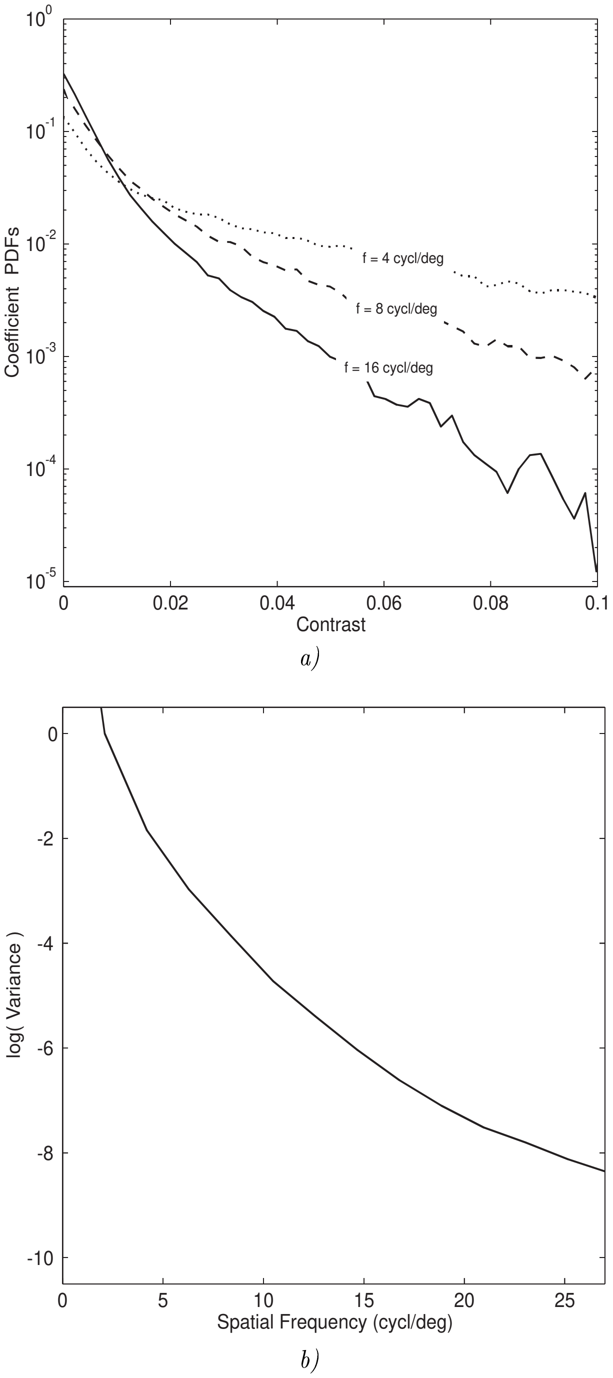


Fig. 1. a) Contrast (normalized amplitude) pdfs for some transform coefficients, b) Variance of the coefficients in the training set.

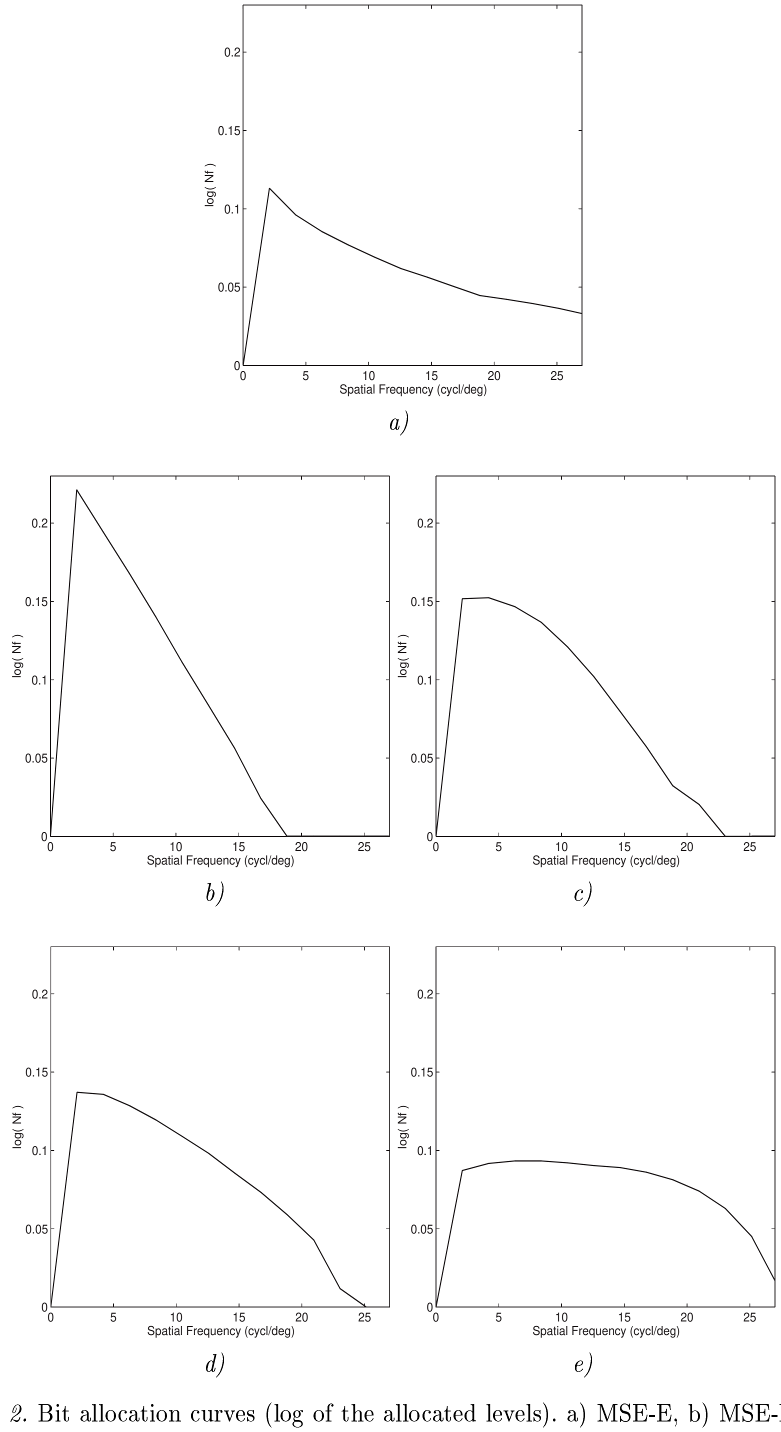
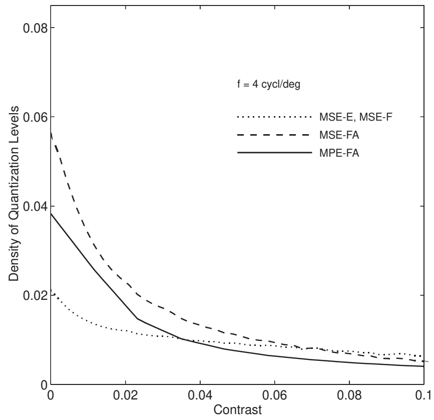
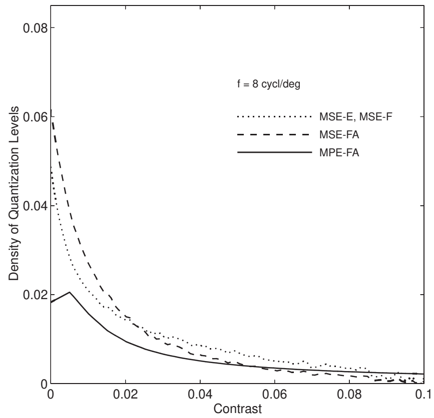


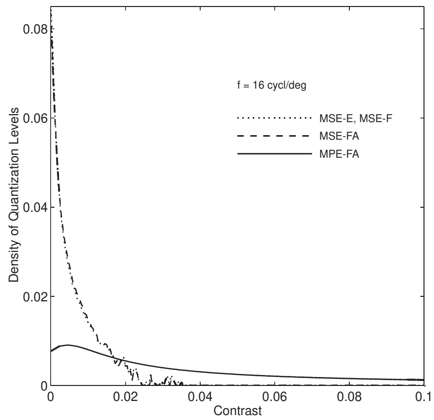
Fig. 2. Bit allocation curves (log of the allocated levels). a) MSE-E, b) MSE-F, c) MSE-FA, d) MPE-F, e) MPE-FA. The curves are set to zero for the first coefficient, the DC component, because it is DPCM coded in every approach.



a)



b)



c)

Fig. 3. Densities of quantization levels, $\lambda_f(a_f)$, for different coefficients using different design criteria. a) 4 cycl/deg, b) 8 cycl/deg, c) 16 cycl/deg.

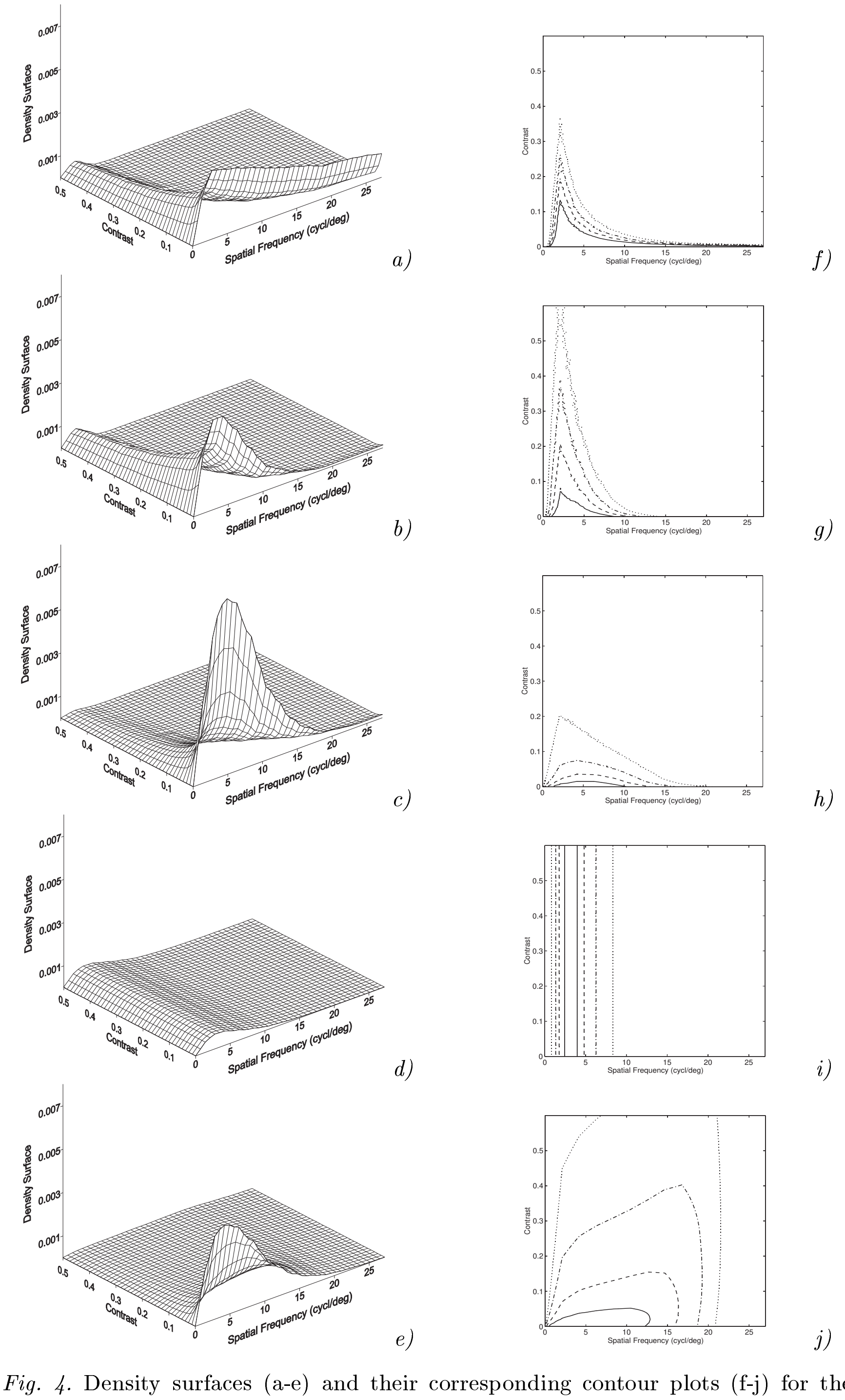


Fig. 4. Density surfaces (a-e) and their corresponding contour plots (f-j) for the different approaches. From top to bottom, MSE-E, MSE-F, MSE-FA, MPE-F and MPE-FA. The contour lines enclose the encoding regions with the 80%, 60%, 40% and 20% of the quantization levels. They show where the encoding effort is focused in the different approaches.

Fig. 5. *Compression results on Barbara image (0.5 bpp). The figures show the* 170×170 *central subimage.*

Fig. 6. *Compression results on Lena image (0.5 bpp). The figures show the* 170×170 *central subimage.*

A.4 Bit allocation algorithm for codebook design in vector quantization fully based on HVS nonlinearities for suprathreshold contrasts

Electronics Letters. Vol. 31, N. 15, pp. 1222–1224 (1995)

Filter example: We consider the three-stage filter shown in Fig. 2, which directly implements our approach. To illustrate the detection and diagnosis capability of the approach, we will consider a parametric fault; namely, a deviation in R_1 of 20%. Fig. 3 shows the simulation results for each stage in the test mode. Signals labelled '_NF' correspond to the fault-free circuit and those labelled '_FR1' correspond to the faulty circuit. The input (V_{in}) used is a square signal between 0 and 1V. As can be seen, the third (Fig. 3*a*) and second (Fig. 3*b*) stages operate as expected. The obtained responses are close to the corresponding fault-free output (both signals are indistinguishable). However, in the test of the first stage (Fig. 3*c*), a disparity with the fault-free response is observed; namely, a fault in the DC gain. Using this measurement and interpreting the first-order transfer function corresponding to the faulty stage, we can predict the cause of the fault. R_1, R_2 and R_7 are responsible for the DC gain, but R_2, R_7 and C_2 also determine the time constant of the stage response. Since this time constant is as expected (as shown in Fig. 3), the error is in R_1. Hence, the fault is detected and its diagnosis coincides with the real fault.

Conclusions: A DfT methodology has been reported for analogue fault diagnosis in active analogue filters. This methodology improves that reported in [8, 9], and its impact on the filter performance is negligible. In terms of hardware and software, the cost is low, requiring only opamps with a duplicated input stage and a previous simulation to obtain the nonfaulty output signals for each stage in the test mode. The applicability has been demonstrated with a simple example using HSPICE simulations. Since each stage of the filter is tested separately, single-faults and multiple-faults can be diagnosed. The methodology can be extended to other types of filters and circuits using opamps.

Acknowledgement: The authors wish to thank A. H. Bratt for his valuable suggestions to improve this work. This work is part of the AMATIST ESPRIT-III Basic Research Project, funded by the CEC under contract 8820.

12 May 1995
Electronics Letters Online No: 19950883

D. Vázquez, A. Rueda and J.L. Huertas (*Dpto. de Diseño de Circuitos Analógicos, Centro Nacional de Microelectrónica, Universidad de Sevilla, Edificio CICA, Avda Reina Mercedes s/n, E-41012, Sevilla, Spain*)

A.M.D. Richardson (*Microelectronics Group, SECAMS, Lancaster University, LA1 4YR, United Kingdom*)

References

1 FASANG, P.P., MULLINS, D., and WONG, T.: 'Design for testability for mixed analog/digital ASICs'. Proc. IEEE 1988 Custom Integr. Circ. Conf., 1988, pp. 16.5.1–16.5.4

2 WAGNER, K.D., and WILLIAMS, T.W.: 'Design for testability of mixed-signal integrated circuits'. Proc. IEEE 1988 Int. Test Conf., 1988, pp. 823–828

3 WEY, C.L.: 'Built-in self-test (BIST) structure for analog circuit fault diagnosis', *IEEE Trans. Instrum. Meas.*, 1990, **39**, pp. 517–521

4 SCHAFER, G., SAPOTTA, H., and DENNER, W.: 'Block-oriented test strategy for analog circuits'. Proc. ESSCIRC, 1991, pp. 217–220

5 HUERTAS, J.L., RUEDA, A., and VAZQUEZ, D.: 'Improving the testability of switched-capacitor filters', *J. Electron.Test., Theory Appl.*, 1993, **4**, (4), pp. 299–313

6 VAZQUEZ, D., HUERTAS, J.L., and RUEDA, A.: 'A new strategy for testing analog filters'. Proc. IEEE VLSI Test Symp., April 1994, pp. 36–41

7 OHLETZ, M.: 'Hybrid built-in self-test (HBIST) for mixed analogue/digital integrated circuits'. Proc. Europ. Test Conf., 1991, pp. 307–316

8 SOMA, M.: 'A design-for-test methodology for active analog filters'. Proc. IEEE Int. Test Conf., 1990, pp. 183–192

9 SOMA, M., and KOLARIK, V.: 'A design-for-test technique for switched-capacitor filters'. Proc. IEEE VLSI Test Symp., 1994, pp. 42–47

10 BRATT, A.H., HARVEY, R.J., DOREY, A.P., and RICHARDSON, A.M.D.: 'Design-for-test structure to facilitate test vector application with low performance loss in nontest mode', *Electron. Lett.*, 1993, **29**, (16), pp. 1438–1440

Bit allocation algorithm for codebook design in vector quantisation fully based on human visual system nonlinearities for suprathreshold contrasts

J. Malo, A.M. Pons and J.M. Artigas

Indexing terms: Image processing, Vector quantisation, Human factors

A fully perceptual oriented approach for codebook design in the vector quantisation of images for compression purposes is presented. The proposed algorithm not only exploits the usually considered dependence on frequency of the human eye, but also makes use of the nonlinear behaviour of the visual system for suprathreshold contrasts. From experimental data of human contrast incremental thresholds of sinusoidal grids, we present a perceptually optimal information allocation function. Simulations show that by introducing perceptual contrast nonuniformities, a more efficient bit allocation is attained than that for the JPEG-like contrast uniform algorithm.

Introduction: The central aim of lossy image compression techniques is to adapt redundancy reduction algorithms to the limitations of the human visual system (HVS), to make the variations unnoticeable. The more recent and successful lossy compression methods are those that perform a vector quantisation (VQ) of the coefficients of a given image transform [1]. The main problem to overcome with these VQ algorithms is the design of a codebook that optimises the subjective quality of the decoded images. A variety of VQ compression algorithms base their codebook design on statistical properties of the images [2]. However, the effect of quantisation on the human eye may not always correspond exactly with mathematical differences in error levels. The HVS capabilities have been introduced as a coefficient-dependent weighting function [3]. In these perceptually weighted algorithms, the number of quantisation levels per coefficient is determined by the human threshold sensitivity whereas the amplitude nonuniformities are still given by the statistical properties of the training set [2], or are not considered [4].

We present an information allocation algorithm for codebook design that can give rise to a general nonuniform partition of the coefficient-amplitude transform domain and can be easily related to the perceptual capabilities of the HVS in suprathreshold conditions. This algorithm improves the subjective quality of the decoded images.

Codebook design controlled by information allocation function (IAF): VQ of a transformed signal involves partitioning of the coefficient-amplitude space. in the proposed scheme, the whole codebook design process is fixed by the information allocation function (IAF). The IAF is a function defined in the coefficient-amplitude space giving the amount of information assigned to encode each zone of that space, given by

$$IAF(f,C) = \frac{d^2 I}{df\,dC} \qquad (1)$$

where I is the information (in bits), and f and C indicate the zone of the coefficient-amplitude (frequency-contrast) space.

The central idea is that the nonuniform distribution of codewords in the vector space must be derived from the IAF in such a manner that we have a peak density of codewords (and minimum values of quantisation step size) where the IAF reaches its maximum. The number of bits assigned to encode each coefficient f can be immediatly obtained by integrating the IAF for amplitudes. We call this integral the cumulative information allocation function (CIAF), where

$$CIAF(f) = \frac{dI}{df} = \int_{C_m}^{C_M} IAF(f,C)dC \qquad (2)$$

where C_m and C_M are the boundaries of the amplitude range.

The CIAF gives us a criterion to make a coefficient-selective bit allocation, but gives us no information about the distribution of the quantisation levels over the amplitude range of a given coefficient. This question is solved by using a set of look-up tables (one per coefficient) called nonuniform quantisation functions (NUQFs) obtained from the IAF by the following equation:

$$NUQF(f,C) = \frac{C_M - C_T(f)}{\int_{C_m}^{C_M} \frac{1}{IAF(f,C)} dC} \int_{C_m}^{C} \frac{1}{IAF(f,C)} dC + C_T(f) \tag{3}$$

where $C_T(f)$ is the threshold amplitude for the coefficient f. The NUQF are monotonically increasing functions representing a nonlinear mapping of the amplitude range $[C_m, C_M]$ into the range $[C_T(f), C_M]$. The slope of the NUQF varies with the amplitude nonuniformities of the IAF, having a minimum slope when the IAF reaches a maximum. To summarise: the IAF represents an intuitive distribution of information in the space to be partitioned, the CIAF gives the number of quantisation levels assigned to each coefficient and each NUQF gives the definitive nonuniform spacing of those quantisation levels. Fig. 1 summarises (with a 1-D example) the codebook design process controlled by the IAF.

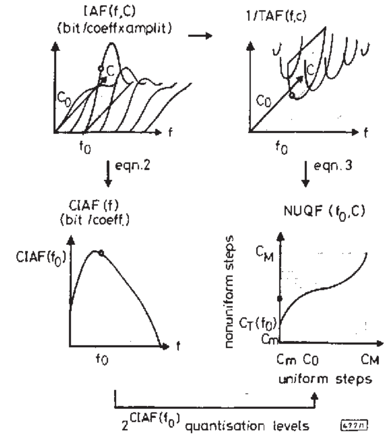


Fig. 1 *1-D example of codebook design process controlled by IAF*

Perceptually optimal information allocation function: In applications where the image is going to be judged by a human observer, VQ must benefit from the visual system limitations, reducing information where the system is not sensitive. In the most general case, the sensitivity peaks of the system are not overlapped with the probability density peaks. Therefore the subjective rule can be stated as follows: the bigger the sensitivity is, the narrower the quantisation step must be. The implementation will be optimal if the quantisation step widths equal (or are proportional) to the HVS tolerances to changes in the coefficient or amplitude, for every point in the coefficient-amplitude space.

In the psychophysical perception literature, those limitations of the ability of discrimination are referred to as incremental thresholds of frequency (Δf) and contrast (ΔC) [5, 6]. If the variation of the coefficients (or amplitude) in the encoding process is lower than the corresponding incremental threshold Δf (or ΔC), the system cannot perceive the variation.

The IAF codebook design algorithm can be easily adapted to obtain quantisation step sizes proportional to the system tolerances as required for optimal subjective performance. A perceptually optimal bit allocation for spatial frequency encoding can be obtained by defining the IAF as the inverse of the area of psychovisual tolerance regions. That is

$$IAF(f,C) = \frac{K}{\Delta f \Delta C} \tag{4}$$

where K is a normalisation constant given by the total amount of information employed when encoding the image. Eqn. 4 states that more information is assigned to the narrower tolerance regions. Recent exhaustive experimental measurements have shown that tolerance of the system to changes in the coefficient or in the amplitude of each basic stimulus depends on the point of the coefficient-amplitude space [5, 6]:

$$\Delta f(f,C) = af + b \tag{5a}$$

$$\Delta C(f,C) = C_T(f) + \frac{C}{\left(\frac{C_T(f)}{k(F)}\right)^{\frac{1}{n(f)}} + C}\left(k(f)C^{n(f)} - C_T(f)\right) \tag{5b}$$

where a and b are constants [5], $n(f)$ and $k(f)$ are experimentally measured parameters [6] and $C_T(f)$ is the threshold contrast of a sinusoid of frequency f to be just perceived. The inverse of $C_T(f)$ is known as the contrast sensitivity function (CSF) [7] and is usually taken as a reference for frequency-selective functions in perceptually weigthed algorithms.

Taking into account the slow variation of the frequency tolerance ($a = 0.025$ for f in cycles per degree[5]) and considering that discrete implementation of the transforms in the encoding process implies a fixed uniform frequency sampling and avoids frequency variations, we will consider Δf = constant. Therefore we propose a perceptually optimal IAF given by the inverse of eqn. 5*b*.

The suprathreshold term in the IAF (the C dependent term) represents a significant qualitative improvement compared with previous perceptually weighted algorithms [4]. If such a suprathreshold dependence were removed, the bit allocation would be controlled by the CSF (the human frequency selective function) as in the standard perceptual oriented VQ methods. Therefore, for example, the JPEG compression algorithm can be derived from our approach for the particular case of contrast independent IAF. The suprathreshold nonuniformities give a physiological basis to empirical changes needed in the quantisation matrix of the JPEG standard when exclusively supported by the threshold sensitivity (filter function CSF).

a *b*

Fig. 2 *Image processed by uniform JPEG-like method (1.6 bit/pixel) and perceptual IAF nonuniform method (1.6 bit/pixel)*

a Uniform JPEG-like method
b Perceptual IAF nonuniform method

Results: We compare our perceptual contrast nonuniform bit allocation algorithm with the JPEG-like algorithm (uniform step size over the entire contrast range). The same number of quantisation levels per coefficient are used in both schemes. Fig. 2*a* and *b* show decoded *Lena* images obtained by applying each algorithm. Imposing the same compression ratio, simulations clearly show a better subjective quality for the contrast nonuniform scheme.

Despite the relatively low compression ratio of the proposed example, the results derived from simulations demonstrate the importance of considering the HVS nonlinearities for suprathreshold contrast in future research.

6 March 1995
Electronics Letters Online No: 19950863

J. Malo, A.M. Pons and J.M. Artigas (*Departament Interuniversitari d'Optica, Facultat de Fisica, Universitat de València, C/Dr. Moliner 50, 46100 Burjassot, València, Spain*)

References

1 GERSHO, A., and GRAY, R.M.: 'Vector quantisation and signal compression' (Kluwer Academic Press, 1992)
2 LINDE, Y., BUZO, A., and GRAY, R.M.: 'An algorithm for vector quantizer design', *IEEE Trans.*, 1980, **COM-28**, pp. 84–95
3 MACQ, B., and SHI, H.Q.: 'Perceptually weighted vector quantization in the DCT domain', *Electron. Lett.*, 1993, **29**, (15), pp. 1382–1384

4 WALLACE, G.K., OMACHI, T., and LEGER, A.: 'JPEG still picture compression algorithm', *Opt. Eng.*, 1991, **30**, (7), pp. 947–954

5 GREENLEE, M.H., and THOMAS, J.P.: 'Effect of pattern adaptation on spatial frequency discrimination', *J. Opt. Soc. Am. A*, 1992, **9**, pp. 857–862

6 PONS, A.M.: 'Spatial properties of contrast discrimination function'. PhD Thesis, Departament Optica, Universitat de Valéncia, 1993

7 KELLY, D.H.: 'Receptive-like functions inferred from large area psychophysical measurements', *Vis. Res.*, 1985, **25**, (12), pp. 1895–1900

Counting lattice points on ellipsoids: Application to image coding

J.M. Moureaux, M. Antonini and M. Barlaud

Indexing terms: Vector quantisation, Image coding

The enumeration problem of lattice vectors in a lattice codebook shaped for elliptical source statistics is solved. The counting of codebook vectors is crucial for the computation of the bit rate, i.e. for the determining the codeword lengths to be transmitted to the decoder.

Introduction: In lattice-based vector quantisation, it is known that the shape of truncation must fit, as closely as possible, the probability density function of the source to minimise the overload noise of the quantiser. This is strongly linked to the shaping advantage of vector quantisation over scalar quantisation, pointed out by Lookabaugh *et al.* in [6]. This gain is obtained by taking into account the spatial distribution of the source vectors in the design of the codebook. Thus, uniform, Gaussian, and Laplacian memoryless sources lead to hypercubic, spherical, and pyramidal codebooks, respectively.

Gaussian sources with memory are also of great interest in signal processing, particularly in image or speech coding. For example, wavelet coefficients [1] of robot vision or satellite images are strongly correlated. Thus, vectors constituted by neighbouring coefficients are elliptically distributed around zero in a privileged direction parallel to $(1, ..., 1)$. Lattice vector quantisation applied to these data will be more efficient when applying an elliptical truncation [2, 5].

We address the case of Gaussian sources with memory, coded by elliptical lattice-based codebooks. In a prefix code encoding scheme, such codebooks need to solve the problem of counting lattice points lying on elllpsoids for codeword length computation [7]. Indeed, knowledge of the total number of codebook vectors [2] is not sufficient in such an encoding scheme. We propose a solution based on the generating theta series designed by Conway and Sloane [3] for the main lattices $\mathbf{Z}^n$, D_n, E_8 which are widely used in data compression applications.

Lattices and elliptical surfaces: The problem of enumeration can be solved by defining concentric surfaces of constant radius on which lattice points can be counted. Thus, the generating theta series proposed by Conway and Sloane enable the counting of lattice points lying on spheres and therefore within spherical codebooks. Furthermore, Fischer proposed a solution for pyramidal codebooks designed in the $\mathbf{Z}^n$ lattice [4] and Solé proposed a solution for pyramidal codebooks designed in the main lattices [8]. We propose to solve the problem for elliptical codebooks designed in the main lattices. First, we can define a centred n-dimensional ellipsoid with axis lengths equal to $a_i r$ as the set

$$\{X \in \mathbf{R}^n | (\|X\|_2^2)_W = r^2\} \tag{1}$$

where $(\|x\|_2^2)_W = XWX^t$, r is a nonnegative constant and W is a weighted diagonal matrix such that $w_{ii} = 1/a_i^2$, $a_i > 0\ \forall\ i$ and $w_{ij} = 0\ \forall\ i \neq j$. Thus, the counting of lattice vectors lying within an elliptical codebook such as $r = l$ can be seen as the counting of lattice vectors lying on concentric ellipsoids for r varying from 0 to l.

Modified theta series for counting lattice points on ellipsoids: For a given lattice Λ, the number N_r of points at distance r from the origin (metric in the weighted L_2 norm) is given by

$$\begin{aligned}\Theta_\Lambda^W(q) &= \sum_{X\in\Lambda} q^{(\|X\|_2^2)_W} \\ &= \sum_{r=0}^{+\infty} q^{r^2}\, |\{X \in \Lambda | (\|X\|_2^2)_W = r^2\}|\end{aligned} \tag{2}$$

The coefficient of q^2 is the number of lattice vectors lying on a centred ellipsoid of radius r (metric in the weighted L_2 norm). Conway and Sloane solved the problem for $W = I$ where I is the identity matrix and propose a method of counting lattice points lying on spheres (using theta series). For more details see [3]. We propose to solve the problem for $W \neq I$ and thus to count lattice points lying on ellipsoids. We first recall the Jacobi theta functions

$$\theta_2(q) = \sum_{m=-\infty}^{+\infty} q^{(m+1/2)^2} \qquad \theta_3(q) = \sum_{m=-\infty}^{+\infty} q^{m^2}$$

$$\theta_4(q) = \sum_{m=-\infty}^{+\infty} (-q)^{m^2} \tag{3}$$

where $q = e^{i\pi z}$. We note that

$$\Theta_{\mathbf{Z}}(q) = \theta_3(q) \qquad \theta_4(q) = \theta_3(q^4) - \theta_2(q^4)$$
$$\text{and } \theta_3(q) = \theta_3(q^4) + \theta_2(q^4) \tag{4}$$

It is useful for computing theta series to consider lattices constructed from algebraic codes. Conway and Sloane give two constructions based on binary linear codes to design main lattices: construction A and construction B [3]. Furthermore, to each binary linear code C is attached a weight enumerator $W_C(x, y)$, where x and y denote the number of zeros and ones in a codeword, respectively. The weight enumerators of codes associated with the main lattices are well known. Furthermore, two important theorems (Theorems 3 and 15, Chapter 7 of [3]) enable us to write

$$\Theta_\Lambda(q) = W_C(\theta_3(q^4), \theta_2(q^4)) \tag{5a}$$
for construction A
$$\Theta_\Lambda(q) = \frac{1}{2}W_C(\theta_3(q^4), \theta_2(q^4)) + \frac{1}{2}[\theta_4(q^4)]^n \tag{5b}$$
for construction B

From the Jacobi functions given in eqn. 3, we can define

$$\theta_j^{a_i}(q) = \theta_j(q^{1/a_i^2}) \tag{6}$$

where $q = e^{i\pi z}$ and $a_i \neq 0$. Then, using construction A for $\mathbf{Z}^n$ and D_n lattices, and construction B normalised by 1/2 for the E_8 lattice, the corresponding weight enumerators are given by

$$\begin{aligned} W_{C(\mathbf{Z}^n)}(x,y) &= \prod_{i=1}^{n}(x_i + y_i) \\ W_{C(D_n)}(x,y) &= \prod_{i=1}^{n}[(x_i + y_i)] + \prod_{i=1}^{n}[(x_i - y_i)] \\ W_{C(2E_8)}(x,y) &= \prod_{i=1}^{8} x_i + \prod_{i=1}^{8} y_i \end{aligned} \tag{7}$$

From eqns. 5, 7 and 4, we can define modified theta series in terms of functions given by eqn. 6 as

$$\begin{aligned} \Theta_{\mathbf{Z}^n}(q) &= \prod_{i=1}^{n}[\theta_3^{a_i}(q^4) + \theta_2^{a_i}(q^4)] = \prod_{i=1}^{n}[\theta_3^{a_i}(q)] \\ \Theta_{D_n}^W(q) &= \frac{1}{2}\left[\prod_{i=1}^{n}[\theta_3^{a_i}(q^4) + \theta_2^{a_i}(q^4)] + \prod_{i=1}^{n}[\theta_3^{a_i}(q^4) - \theta_2^{a_i}(q^4)]\right] \\ &= \frac{1}{2}\left[\prod_{i=1}^{n}\theta_3^{a_i}(q) + \prod_{i=1}^{n}\theta_4^{a_i}(q)\right] \\ \Theta_{2E_8}^W(q) &= \frac{1}{2}\left[\prod_{i=1}^{8}\theta_2^{a_i}(q^4) + \prod_{i=1}^{8}\theta_3^{a_i}(q^4) + \prod_{i=1}^{8}\theta_4^{a_i}(q^4)\right] \Rightarrow \\ \Theta_{E_8}^W(q) &= \frac{1}{2}\left[\prod_{i=1}^{8}\theta_2^{a_i}(q) + \prod_{i=1}^{8}\theta_3^{a_i}(q) + \prod_{i=1}^{8}\theta_4^{a_i}(q)\right] \end{aligned} \tag{8}$$

A.5 Splitting criterion for hierarchical motion estimation based on perceptual coding

Electronics Letters. Vol. 34, N. 6, pp. 541–543 (1998)

encoded at a rate of 10kbit/s. Image quality was evaluated by peak signal-to-noise ratio (PSNR). For comparison, BMA (block matching algorithm) and the conventional 2D triangular mesh based [3] methods were also implemented in our experiments. In BMA, mean absolute difference is used as the prediction error criterion and the search area is up to [−15, +15] pixels in the horizontal and vertical directions around the original 16 × 16 macroblock position. The PSNR and the normalised execution time produced by these three methods are provided in Figs. 1 and 2, respectively.

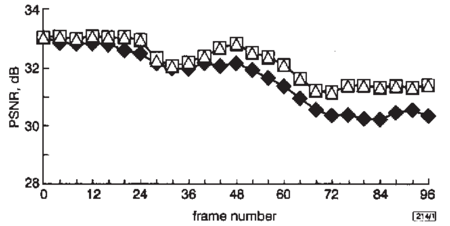


Fig. 1 *PSNR of results for 'Mother and Daughter' at 10kbit/s and 7.5f/s*

◆ BMA
■ 2D mesh
△ proposed

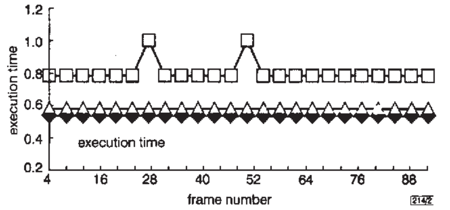


Fig. 2 *Execution time for different methods*

◆ BMA
□ 2D mesh
△ proposed

From Fig. 1, it is observed that the PSNR performance of the proposed method is better than that of the BMA by a factor of 0.55dB. Conversely, the PSNR of the proposed method is very close to that of the conventional 2D triangular mesh based method. From Fig. 2, it can also be seen that the processing speed of the proposed method is nearly 37% better than that of the 2D triangular mesh based method, and is close to that produced by the BMA method.

Conclusions: A fast motion estimation method for 2D triangular mesh based image coding has been presented. The proposed method solves two big problems of previous methods. One is the low image quality in the BMA method, and the other is the expensive computational cost in the conventional 2D triangular mesh based method. The simulation results show that the average PSNR of the proposed method is better than that of the BMA method, and the computational cost of the proposed method is cheaper than that of the conventional 2D triangular mesh based method. Therefore, the proposed algorithm can be very useful for very low bit rate real-time video coding applications.

 5 January 1998
Electronics Letters Online no: 19980400

Hak Soo Kim, Taekhyun Yun and Kyu Tae Park (*Computer Application Laboratory, Department of Electronic Engineering, Yonsei University, 134 Shinchon-dong, Seoduemun-gu, Seoul 120-749, Korea*)

E-mail: hskim@eve.yonsei.ac.kr

References

1 NAKAYA, Y., and HARASHIMA, H.: 'An iterative motion estimation method using triangular patches for motion compensation'. SPIE Visual Commun. and Image Processing '91: Visual Commun., Boston, MA, November 1991, Vol. 1605, pp. 546–557

2 CHOI, C.S., AIZAWA, K., HARASHIMA, H., and TAKEBE, T.: 'Analysis and synthesis of facial image sequences in model-based image coding', *IEEE Trans. Video Technol.*, 1994, **4**, pp. 257–275

3 NAKAYA, Y., and HARASHIMA, H.: 'Motion compensation based on spatial transforms', *IEEE Trans. Video Technol.*, 1994, **4**, pp. 339–356

4 ROSS, S.M.: 'Introduction to probability and statistics for engineers and scientists' (John Wiley & Sons, 1987)

Splitting criterion for hierarchical motion estimation based on perceptual coding

J. Malo, F.J. Ferri, J. Albert and J.M. Artigas

A new entropy-constrained motion estimation scheme using variable-size block matching is proposed. It is known that fixed-size block matching as used in most video codec standards is improved by using a multiresolution or multigrid approach. In this work, it is shown that further improvement is possible in terms of both the final bit rate achieved and the robustness of the predicted motion field if perceptual coding is taken into account in the motion estimation phase. The proposed scheme is compared against other variable- and fixed-size block matching algorithms.

Introduction: Motion estimation is a crucial step in video coding, mainly because optical flow is used to remove temporal redundancy from the signal. In current standards, the fixed-size block matching algorithm (BMA) is commonly used for motion compensation due to its trade-off of simplicity, computational load and performance [1].

Nevertheless, a fixed-size BMA usually implies a bad trade-off between motion information and prediction error volume because many motion vectors are wasted in stationary areas while more precision is needed along moving edges. Also, it exhibits a certain tendency towards unstable estimates [2].

To partially overcome these drawbacks, several hierarchical or variable-size BMAs have been proposed [2 – 5]. By operating in a top-down manner at different resolution levels they can give reliable, locally adapted motion estimates. Apart from different search strategies and similarity measures used, the splitting criterion is the most important part of most multigrid schemes because it allows the algorithm to locally refine the motion field. Several splitting criteria have already been used:

(i) A measure of the magnitude of the prediction error (MSE or MAE) [3, 4].

(ii) A measure of the volume of the signal to be sent to the decoder. The entropy of the motion field plus the zero-order entropy of the error signal in the spatial domain has been used [2, 5].

It has been pointed out that the entropy-based criterion is better than other criteria based on frame differences because, by its nature, it is well suited for coding applications. Moreover, it does not require any pre-specified threshold as in the first criterion.

As all of the above criteria use spatial domain measures, it cannot be guaranteed that the splitting process will be stopped when no further coding benefit is obtained. Moreover, this over-resolution may give rise to unstable motion estimates and false alarms. In fact, the need for a merging step has been pointed out to avoid this problem [3].

New splitting criterion: The use of entropy-based measures adapts the motion prediction to the particular goal of coding, but it should be emphasised that reducing the volume of the error signal in the spatial domain [2, 5] does not exactly match the behaviour of most image coding schemes in which the final step is a frequency-dependent quantisation. The actual volume of the prediction error is the entropy of its quantised transform coefficients (*spectral entropy*) and not the entropy of the pixels of the error (*spatial entropy*). In this Letter, a new splitting criterion based on spectral entropy is proposed, in the context of entropy-constrained variable-size BMA, in the following way:

Split a block at a given resolution in a multigrid scheme if:

$$H\left[\vec{D}_{split}\right]+H_f[DFD_{split}] < H\left[\vec{D}_{nosplit}\right]+H_f[DFD_{nosplit}]$$

where $H(\vec{D})$ is the entropy of the displacement vectors and H_f (DFD) stands for spectral entropy of the displaced frame difference (DFD), which is defined as the entropy of the quantised error signal in the DCT domain:

$$H_f[DFD] = H[Q[DCT[DFD(\mathrm{x})](\mathrm{f})]]$$

This change really makes a difference because splitting a block using previous criteria will give rise to a reduction in complexity of the error signal in the spatial domain that may be completely ignored by the quantiser which usually carries out frequency perceptual information [1, 6].

If this perceptual information is taken into account in the splitting criterion (using the quantised signal), it is possible to obtain a more compact motion description and a better trade-off between motion information and prediction error: first, because the splitting criterion is more closely connected to the way in which the coder performs, thus increasing the precision only if the corresponding quantiser can take advantage of it; and secondly, because perceptual information gives rise to a motion estimate that implies more accurate compensations from an observer's point of view. This means that perceptually important features of the scene (seen as sharp moving edges) receive more effort for prediction. Even though this scheme obviously requires more computation than spatial-based algorithms, this additional effort is restricted to the (usually off-line) encoding phase while the decoding phase remains unchanged.

Results and discussion: Several experiments have been carried out using different standard image sequences. Fixed- and variable-size BMAs using both spatial and spectral entropy criteria have been considered at the same bit rate (200kbit/s) for comparison purposes. Blocks of size 8×8 have been used in a fixed-size BMA. A quadtree multigrid scheme with five resolution levels (blocks from 64×64 to 4×4) has been used in a variable-size BMA. The *n*-step displacement search [2], has been applied to compute motion vectors at all resolution levels. The usual correlation has been used as a similarity measure. A nonlinear perceptually-weighted DCT quantiser has been used to take into account the frequency and amplitude response of the human viewer [6].

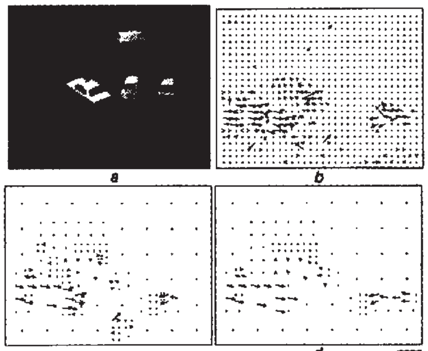


Fig. 1 *Estimated motion field using three BMA considered*

a Frame 6 of *Taxi* sequence
b Fixed-size
c Spatial entropy criterion
d Spectral entropy criterion

Frame six of the *Taxi* sequence is shown in Fig. 1, along with the motion field estimate using the three algorithms considered. It is worth noting that with the proposed criterion, fewer motion vectors are needed to properly describe the motion in the scene. Fewer false alarms are obtained in the static areas, and significant motions are better described. Note how in the static area at the bottom of Fig. 1*c*, the spatial entropy-based algorithm gives the same false alarm as the uniform resolution BMA (Fig. 1*b*), due to an inadequate increase in resolution. The motion of the incoming vehicle on the right is better described by the proposed method.

A side-effect of using perceptual information is a more conservative splitting strategy which gives rise to more robust estimates. This is because perceptual quantisers show bandpass behaviour [1, 6] neglecting certain aspects of the prediction error that would have been considered by the objective spatial domain based criteria. In fact, resolution is increased only if the corresponding motion structures are perceptually relevant. Most moving objects can be properly identified with the proposed criterion, giving rise to coherent fields without spurious vectors, unlike in the other cases.

Table 1: Motion field volumes (kbit/s) for different sequences

	Taxi	Rubik	Yosemite	Trees
Fixed-size BMA	25.15	43.42	92.37	78.42
Spatial entropy	3.78	6.58	17.38	9.94
Spactral entropy	1.87	3.73	8.64	6.48

Table 1 shows the volume of the motion description using the three algorithms. With the proposed criterion, the motion information needed for compensation is approximately reduced by factors of 2 and 10 with regard to spatial entropy-based variable-size BMA and fixed-size BMA, respectively. Savings in motion information can be spent by the new approach in a more accurate encoding of the prediction errors, giving rise to subjectively better reconstructions.

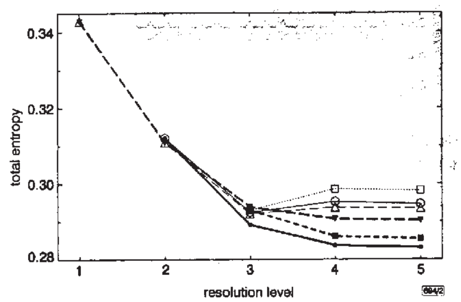


Fig. 2 *Volume of flow and prediction errors (in bit/pix) while refining motion estimate for two variable-size BMA considered at three different starting resolution levels*

▼— —▼ spectral criterion (i)
●——● spectral criterion (ii)
■---■ spectral criterion (iii)
△- - -△ spatial criterion (i)
○——○ spatial criterion (ii)
□·······□ spatial criterion (iii)

At each frame, the entropy of the encoded signal (flow and quantised errors) has been computed at the different resolution levels of the quadtree during the application of the variable-size BMA. Fig. 2 shows these results averaged over all frames. Different starting resolution levels are considered with both splitting criteria. As expected, the proposed criterion does minimise the volume of the signal in the meaningful encoding domain, while the spatial entropy criterion gives rise to an increase in volume during the final steps, presumably because too many blocks are unnecessarily split.

Conclusion: A new splitting criterion for hierarchical BMA has been proposed. From the viewpoint of video coding, the new method improves on other approaches because the behaviour of the quantiser is taken into account for motion estimation. In this way, motion estimation effort is directed towards certain areas in such a way that the final bit rate is minimised. Conversely, due to perceptual weighting, the final flow exhibits some interesting properties, such as coherence, compactness, robustness and reliability, that could make it suitable for more general applications.

Acknowledgment: This work was partially supported by the Spanish CICYT project TIC95-676-C02-01.

27 November 1997
Electronics Letters Online No: 19980404

J. Malo and J.M. Artigas (*Dept. d'Òptica, Universitat de València, Dr. Moliner, 50, 46100 Burjassot, Spain*)

F.J. Ferri and J. Albert (*Dept. d'Informàtica i Electrònica, Universitat de València, Dr. Moliner, 50, 46100 Burjassot, Spain*)

E-mail: ferri@uv.es

References

1 LEGALL, D.: 'MPEG: a video compression standard for multimedia applications', *Commun. ACM*, 1991, **34**, (4), pp. 47–58

2 DUFAUX, F., and MOSCHENI, F.: 'Motion estimation techniques for digital TV: A review and a new contribution', *Proc. IEEE*, 1995, **83**, (6), pp. 858–875

3 CHAN, M.H., YU, Y.B., and CONSTANTINIDES, A.G.: 'Variable size block matching motion compensation with applications to video coding', *IEE Proc.*, 1990, **137**, (4), pp. 205–212

4 LI, J., and LIN, X.: 'Sequential image coding based on multiresolution tree architecture', *Electron. Lett.*, 1993, **29**, (17), pp. 1545–1547

5 MOSCHENI, F., DUFAUX, F., and NICOLAS, H.: 'Entropy criterion for optimal bit allocation between motion and prediction error information'. SPIE Proc. Visual Communication and Image Processing, 1993, pp. 235–242

6 MALO, J., PONS, A.M., and ARTIGAS, J.M.: 'Bit allocation algorithm for codebook design in vector quantisation fully based on human visual system non-linearities for suprathreshold contrasts', *Electron. Lett.*, 1995, **31**, (15), pp. 1222–1224

Video object motion representation using run-length codes

M.K. Steliaros, G.R. Martin and R.A. Packwood

A method of runlength coding motion vector data for object based video coding is proposed. It is shown that significant improvements in coding efficiency can be gained for certain classes of source material.

Introduction: Evolving object-based video coding standards, such as MPEG-4, permit arbitrary-shaped objects to be encoded and decoded as separate video object planes (VOPs). In MPEG-4, the exact method of producing VOPs from the source imagery is not defined, but it is assumed that 'natural' video objects are represented by shape information in addition to the usual luminance and chrominance components. Shape data is provided as a binary segmentation mask, or a grey scale alpha plane to represent multiple overlaid objects. The MPEG-4 video verification model [1] proposes that motion compensation (MC)-based predictive coding is applied to each VOP. Conventional fixed size block matching (FSBM) motion estimation, of the type originally proposed by Jain and Jain [2], is employed. In this Letter, we propose two improvements: a best vector selection strategy, and an alternative method of coding motion vector data, which results in a considerable improvement in coding efficiency.

Block matching: In MPEG-4 it is proposed that FSBM is used to estimate the motion in presegmented video objects.

The VOP is constructed by placing 16 × 16 pixel macro-blocks over the shape description, and then further extended by one macro-block to cater for movement of the object between frames. A repetitive padding technique is used to fill out those areas which are inside the block structure but outside the object's shape mask.

The padding is designed to feed the best possible input into the next stage of coding, especially when material that did not exist in the reference frame is introduced. The resulting VOP is used as the reference frame for the block matching operation. Motion estimation is performed by matching the 16 × 16 macro-blocks within a prescribed search area in the reference frame. For maximum accuracy, an exhaustive search is used, and matching is conducted to 1/2 pixel precision. There are several options for the matching criterion [3], although that most commonly used is the sum absolute difference (SAD), since it is not as computationally demanding as others and yet achieves similar performance. It provides a measure of the error between a block of size $n \times n$ pixels in the current frame (C) and a block with displacement (x, y) in the reference frame (R) and can be described as

$$SAD(x,y) = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} |C(i,j) - R(i+x, j+y)| \quad (1)$$

where C (i, j) and R (i, j) represent the intensity of the pixel at location (i, j) in the current and reference frames, respectively. Both ITU-T H.263 [4] and MPEG-4 modify the SAD for the zero vector in order to favour it when there is no significant difference between vectors:

$$SAD(0,0) = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} |C(i,j) - R(i,j)| - \left(\frac{N_B}{2} + 1\right) \quad (2)$$

where N_B is the number of pixels in the block that are inside the VOP. These 'best' vectors are subsequently differentially coded, typically using a two-dimensional (2D) prediction strategy described in more detail in the next section.

Even with the zero bias, however, the above definition does not uniquely identify a 'best' choice when multiple vectors produce identical errors. A naive implementation of the exhaustive search algorithm would perform the search in a raster scan order. If each motion vector is denoted by $\mathbf{v}_i$ (x_i, y_i), where x_i, y_i (or $2x_i$, $2y_i$ for half pixel search) are integers, $-r \le x_i$, $y_i < r$ and r is the search radius, then for a raster scan, all $\mathbf{v}_i$ need to be ordered such that:

$$\mathbf{v}_i < \mathbf{v}_{i+1} \Leftrightarrow (y_i < y_{i+1}) \vee ((y_i = y_{i+1}) \wedge (x_i < x_{i+1})) \qquad \forall\, i \in [0, 4r^2 - 1] \quad (3)$$

Since vector quality affects the 2D prediction, choosing the smallest vector in Euclidean terms will improve coding efficiency. The simplest solution in terms of implementation costs is to impose an optimal ordering in the search algorithm which ensures that smaller vectors are tested before larger ones. We therefore propose that the ordering condition has to be modified to:

$$\mathbf{v}_i < \mathbf{v}_{i+1} \Leftrightarrow (|\mathbf{v}_i|^2 < |\mathbf{v}_{i+1}|^2) \vee \Big((|\mathbf{v}_i|^2 = |\mathbf{v}_{i+1}|^2) \wedge ((x_i < x_{i+1}) \vee ((x_i = x_{i+1}) \wedge (y_i < y_{i+1})))\Big) \qquad \forall\, i \in [0, 4r^2 - 1] \quad (4)$$

2D motion vector prediction: In MPEG-4 it is proposed that FSBM is used to perform motion estimation in pre-segmented video objects. The motion vector (MV) x and y components for each macro-block are transmitted as Huffman-type variable length codes (VLC) which depend on the spatial neighbourhood of three previously transmitted motion vectors and are calculated as the median of the vectors immediately to the left, above and to the above right of the block being considered, with special rules applying near the VOP boundaries.

Image sequences which exhibit small changes in motion from block to block result in small MVDs, and thus require few bits of VLC to represent the motion. In the extreme case, when motion vectors are identical, two bits are used for each block. This might arise in a panning shot or a still picture.

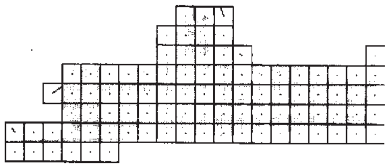


Fig. 1 *FSBM block structure example for 'container ship'*

Fig. 1 shows an example block structure produced by FSBM for a frame from the MPEG-4 test sequence 'container ship'. The short lines shown in the centre of the blocks indicate the motion vector associated with the block. Although most vectors are

A.6 Exploiting perceptual feed-back in multi-grid motion estimation for video coding using an improved DCT quantization scheme

IEEE Transactions on Image Processing (Enviado Abril 1999).

Exploiting perceptual feedback in multigrid motion estimation for video coding using an improved DCT quantization scheme [1]

J. Malo [a], F. J. Ferri [b], J. Soret [b] and J.M. Artigas [a]

[a] *Dpt. d'Òptica*

[b] *Dpt. d'Informàtica i Electrònica*

Universitat de València
C/ Dr. Moliner 50. 46100. Burjassot, València, Spain
e-mail: Jesus.Malo@uv.es

In this work, a multigrid motion compensation video coder scheme based on the current HVS contrast discrimination models is proposed. A novel procedure for the encoding of the prediction errors, which restricts the maximum perceptual distortion in each transform coefficient, has been used. This subjective redundancy removal procedure includes the amplitude non-linearities and some temporal features of human perception. A perceptually weighted control of the adaptive motion estimation algorithm has been derived from this model. Perceptual feed-back in motion estimation ensures a perceptual balance between the motion estimation effort and the redundancy removal process. The results show that this feed-back induces a scale-dependent refinement strategy that gives rise to more robust motion estimates which facilitate higher level sequence interpretation. The comparison of the proposed scheme versus a H.263 scheme with unweighted motion refinement and MPEG-like quantization shows the overall subjective superiority of the proposed improvements.

Key words: Video Coding. Entropy-Constrained Motion Estimation. Perceptual Quantization. Non-linear Perception Model.

EDICS: 1-VIDC Video Coding.

[1] Partially supported by CICYT projects TIC 1FD97-0279 and TIC 98-677-C02-02.

1 Introduction

In natural video sequences to be judged by human observers, two kinds of redundancies can be identified: *objective redundancies*, which are related to the spatio-temporal correlations among the video samples, and *subjective redundancies*, which refer to the data that can be safely discarded without perceptual loss.

The aim of any video coder scheme is to remove both kinds of redundancy from the encoded signal. To accomplish this aim, the video coders currently used are based on motion compensation and 2D transform coding of the residual error [1–3]. The original video signal is split into motion information and prediction errors. These two lower complexity sub-sources of information are usually referred to as Displacement Vector Field (DVF) and Displaced Frame Difference (DFD), respectively.

In the most recent standards, H.263 and MPEG-4 [4, 5], the fixed-resolution motion estimation algorithm used in H.261 and MPEG-1 has been replaced by an adaptive variable-size Block Matching Algorithm (BMA) to obtain improved motion estimates [6]. Spatial subjective redundancy is commonly reduced through a perceptually weighted quantization of a transform of the DFD. The bit allocation among the transform coefficients is based on the spatial frequency response of simple (linear and threshold) perception models [1–3].

In this context, there is a clear trade-off between the effort devoted to motion compensation and transform redundancy removal. On one hand, a better representation of the motion information may lead to better predictions and should alleviate the task of the quantizer. On the other hand, better quantization techniques may be able to succesfully remove more redundancy, thereby reducing the predictive power needed in the motion estimate.

This paper adresses the problem of the trade-off between multigrid motion estimation and error coding using an improved spatio-temporal perception model that takes into account the non-linearities of the Human Visual System (HVS). In this paper, a novel subjective redundancy removal procedure [7, 8], and a novel perceptually weighted motion estimation algorithm [9], are joint together to propose a fully perceptual motion compensated video coder. The role of the perceptual model in the proposed video coder scheme is twofold. First, it is used to define an accurate perceptually matched quantizer in order to obtain better subjective redundancy removal from the error sequence. Second, this quantizer is used to obtain a measure of the perceptually relevant information in the error signal. This perceptual information measure is used to control a multigrid entropy-constrained motion compensation. This control introduces a perceptual feed-back in the motion estimation stage to obtain a subjectively meaningful motion representation and to ensure a perceptual balance between motion information and prediction errors. The particular band-pass shape of the perceptual constraint to the motion estimation gives a scale-dependent control criterion that may be useful for discriminating between significant and noisy motions. Therefore, the benefits of including the properties of the biological filters in a generic video representation scheme go beyond the optimization of the bit rate for a given distortion.

The paper is organized as follows. In Section 2, the current methods of quantizer design for transform coding and variable-block size BMA for motion compensation are briefly reviewed. The proposed improvements in the quantizer design are detailed in section 3 along with their perceptual foundations. In section 4, the proposed multigrid refinement criterion is obtained from the requirement of a monotonic reduction of the significant entropy of DFD and DVF. The results of the proposed improvements are presented in section 5 along with the results of previously reported algorithms. The advantages of the proposed scheme in subjective quality and the interesting qualitative features of its motion estimates are discussed in section 6. To conclude, some final remarks are given in section 7.

2 Conventional techniques for transform quantizer design and multigrid motion estimation

The most important elements of a motion compensated video encoder are the optical flow estimation algorithm and the quantizer. The optical flow provides motion information that leads to a reduction of the objective temporal redundancy, while the quantization of the transformed error signal (usually a 2D DCT) reduces the remaining (objective and subjective) redundancy to a certain extent [1–3].

Signal independent JPEG-like uniform quantizers are employed in the commonly used standards [1, 4]. In this case, the bit allocation in the 2D DCT domain is heuristically based on the threshold detection properties of the HVS [2, 3], but neither amplitude non-linearities [10] nor temporal properties of the HVS [11, 12] are taken into account. The effect of these properties is not negligible [13, 14]. In particular, the non-linearities of the HVS may have significant effects on the bit allocation, improving the subjective results of the JPEG-like quantizers [7, 8, 15].

The conventional design of a generic transform quantizer is based on the minimization of the *average* quantization error over a training set [16]. However, the techniques based on average error minimization have some subjective drawbacks in image coding applications. The optimal quantizers in an average error sense may arbitraryly underperform on individual blocks or frames [17, 18] even if the error measure is perceptually weighted [19]: the accumulation of quantization levels in certain regions in order to minimize the average perceptual error does not ensure good behavior on a particular block of the DFD. This suggests that the subjective problems of the conventional approach are not only due to the use of perceptually unsuitable metrics, as usually claimed, but are also due to the use of an inappropriate *average error* criterion. In addition to this, quantizer designs that depend on the statistics of the input have to be re-computed as the input signal changes.

These factors favour the use of heuristically derived perceptual quantizers in the current standards instead of the conventional, average error-based quantizers.

Multigrid or quadtree-based motion estimation techniques are based on matching between variable-size blocks of consecutive frames of the sequence [6]. The motion estimation starts at a coarse resolution (blocks of large size). The best displacement for each block is computed through a restricted search in each resolution [20]. The resolution of the motion estimate is locally increased (a block of the quadtree segmentation is split) according to some refinement criterion. The motion estimation process ends when no block of the quadtree structure can be split further.

The splitting criterion is the most important part of the algorithm because it controls the local refinement of the motion estimate, which has effects on the relative volumes of DVF and DFD [9, 18, 21, 22], and may give rise to unstable motion estimates due to an excesive refinement of the quadtree structure [9, 23]. This is critical in model-based video coding where an initial adaptive segmentation of the frame can be used as a coarse representation of objects with coherent motion or as a guess for higher level scene interpretation [5, 24, 25]. Two kinds of splitting criteria have already been used:

- A measure of the magnitude of the prediction error [6, 23, 26–29], its energy, mean square error or mean absolute error.
- A measure of the complexity of the prediction error. In this case, spatial based entropy measures [21, 22], or the entropy of the encoded DFD [9, 18, 30, 31] have been reported.

The difference-based criteria were proposed without a specific relation to the quantization process, neglecting the trade off between DVF and DFD that arises from video coding. It has been pointed out that the entropy-based criteria are well suited for coding applications because they minimize the joint volume of the DVF and the DFD [21, 22]. However, it has been shown [9] that the spatial-based entropy measures do not exactly match the behaviour of commonly used coders in which the final stage is a perceptual-based frequency-dependent quantization. The particular band-pass behaviour of the perceptual quantizer may introduce interesting effects into the splitting criterion. Comprehensive rate-distortion-based approaches were designed to obtain the optimal bit allocation between DVF and DFD [18, 30, 31]. However, in spite of the implicit consideration of the DCT coder, the qualitative effects of an uneven quantization of the transform domain were not analyzed.

3 Perceptually uniform DCT quantization

Splitting the original information source into two lower complexity subsources (DVF and DFD) does reduce their redundancy to a certain extent. However, the enabling fact behind very-low-bit-rate is that not all non-redundant data are significant to the human observer. This is why more than just the strictly predictable data can be safely discarded in the lossy quantization process.

According to the current models of human contrast processing and discrimination [32, 33], the input spatial patterns, $\mathbf{A} = \{A_x\}_{x=1}^{n}$, are first mapped onto a local frequency domain through the application of a set of band-pass filters with different relative gains, $\mathbf{a} = T(\mathbf{A})$, with $\mathbf{a} = \{a_f\}_{f=1}^{m}$. After that, a non-linear, log-like transform is applied to the response of each filter, a_f, to obtain the response representation, $\mathbf{r} = R(\mathbf{a})$, with $\mathbf{r} = \{r_f\}_{f=1}^{m}$.

Two facts have favoured the interpretation of the contrast perception at this level as a sort of transform coding process characterized by a pair, (T, Q_p), of a perceptual transform followed by a perceptual quantizer. First, the non-uniformity of the just noticeable differences (JNDs) of the amplitude of the basis functions of the transform T [10, 12, 32, 34] implies a non-euclidean perceptual metric in the transform domain [35, 36]. Second, the HVS cannot discriminate between two amplitude values of the transform basis that differ less than a JND, so the same subjective perception is obtained from both inputs. This has been interpreted as a non-uniform quantization of the signal in the perceptual transform domain [7, 35, 37–39]. The similarity between the impulse responses of the perceptual filters of the transform T and the basis functions of the local frequency transforms used in image and video coding has been used to apply the experimental properties of the perceptual transform domain to the block DCT transform as a reasonable approximation [7, 8, 15, 40–42].

In this paper, a fully perceptual scheme for video compression is proposed from the point of view of preserving no more than the subjectively relevant information. The basic idea is to simulate the redundancy removal in the HVS through an appropriate perceptually inspired quantizer, and then, to use this quantizer to control the adaptive motion estimation. In this way, the motion estimation effort is limited by the requirements of the perceptual quantizer, avoiding the vain prediction of details that are going to be discarded by the quantizer. Here, the perceptual quantization is formulated in the DCT domain through an explicit design criterion based on a distortion metric that includes the HVS non-linearities [35, 36] and some temporal perceptual features [11, 12, 34].

3.1 The Maximum Perceptual Error *criterion for DCT quantizer design*

The natural way of assessing the quality of an encoded picture involves a one-to-one comparison between the original and the encoded version of the image (or short sequence of images). The result of this comparison is related to the ability of the observer to notice the particular quantization noise in presence of the original (masking) pattern. This one-to-one noise detection or assessment is clearly related to the tasks behind the standard pattern discrimination models [32, 33], in which an observer has to evaluate the perceptual distortion in some direction from a masking stimulus.

In contrast, a hypothetical request of assessing the global performance of a quantizer over a set of images or sequences would involve a sort of averaging of each one-to-one comparison. It is unclear how a human observer does this kind of averaging to obtain a global feeling of performance, and the task itself is far from the natural one-to-one comparison that arises when one looks at a particular picture.

The conventional techniques of transform quantizer design use average design criteria in such a way that the final quantizer achieves the minimum average error over the training set (sum of the one-to-one distortions weighted by their probability) [16]. However, the minimization of an average error measure does not guarantee a satisfactory subjective performance on individual comparisons [17, 18]. Even if a perceptual weighting is used, the statistical factors introduced by the average criteria may bias the results. For instance, Macq [19] used uniform quantizers instead of the optimal Lloyd-Max quantizers [16, 43] due to the perceptual artifacts caused by the outliers on individual images.

It has been shown that in order to prevent high perceptual distortions on individual images arising from misbehaved coefficients, the coder should restrict the *maximum perceptual error* in each coefficient and amplitude [7, 8]. This requirement is satisfied with a perceptually uniform distribution of the available quantization levels in the transform domain: if the perceptual distance between levels is constant, the maximum perceptual error in each component is bounded regardless of the amplitude of the input.

In this paper, the restriction of the *maximum perceptual error* (MPE) will be used as a design criterion. This criterion can be seen as a perceptual version of the minimum maximum error criterion [17, 18]. This idea has been implicitly used in still image compression [15, 42] to achieve a constant error contribution from each frequency component on an individual image. It has been shown that bounding the perceptual distortion in each DCT coefficient may be subjectively more effective than minimizing the average perceptual error [7, 8]. The MPE quantizers simplify to the CSF-based uniform quantizers of the JPEG and MPEG schemes if the amplitude non-linearities of the adopted perceptual model are neglected.

3.2 Optimal spatial quantizers under the MPE criterion

The design of a transform quantizer for a given block transform, T, involves finding an optimal distribution of the quantization levels for each transform coefficent and establishing an optimal bit allocation among the transform coefficients [16].

The optimal MPE quantizer design is a function of the subjective metric of the DCT domain. Assuming that the squared perceptual distance between two close patterns in the transform domain, $\mathbf{a}$ and $\mathbf{a}+\Delta\mathbf{a}$, is given by:

$$D^2(\mathbf{a}, \mathbf{a}+\Delta\mathbf{a}) \;=\; \sum_{f=1}^{m} D_f^2 \;=\; \sum_{f=1}^{m} W_f(a_f)\, \Delta a_f^2 \qquad (1)$$

where $W_f(a_f)$ is a generic frequency and amplitude dependent metric, the number of quantization levels per coefficient, N_f, and the density of these levels in amplitude, $\lambda_f(a_f)$, that make the subjective distance between levels constant are [7, 8]:

$$N_f = \frac{1}{2k}\int W_f(a_f)^{1/2}\, da_f \tag{2}$$

$$\lambda_f(a_f) = \frac{W_f(a_f)^{1/2}}{\int W_f(a_f)^{1/2} da_f} \tag{3}$$

where k is a constant depending on the total number of bits used to encode each transform block.

Here, a particular perceptual metric related to the gradient of the non-linear response, R, and to the amplitude JNDs will be used [35, 36]:

$$W_f(a_f) = \left(\frac{\partial R(\mathbf{a})}{\partial a_f}\right)^2 \propto JND(a_f)^{-2} = \left(CSF_f{}^{-1} + a_f\, G_f(a_f)\right)^{-2} \tag{4}$$

where the *Contrast Sensitivity Function*, CSF_f, is the band-pass linear filter which characterizes the HVS performance for low amplitudes [11, 14, 19, 44], and $G_f(a_f)$ are empirical monotonically increasing functions of amplitude for each spatial frequency to fit the amplitude JND data [35].

From this MPE formulation, two interesting cases can be derived. First, if a simple linear perception model is assumed, a CSF-based MPEG-like quantizer is obtained. If the non-linear amplitude-dependent correction in Eq. 4 is neglected, uniform quantizers are obtained for each coefficient, and N_f becomes proportional to the CSF; which is one of the recommended options in the JPEG and MPEG standards [1–3]. Second, if both frequency and amplitude factors of the metric are taken into account, the perceptual algorithm of ref. [15] is obtained: the quantization step size is input-dependent and proportional to the JNDs; and the bit allocation is proportional to the integral of the inverse of the JNDs.

The uniform quantizers used in MPEG [1–3] and the proposed MPE quantizers that use eqs. 2 to 4 [7, 8, 15], represent different degrees of approximation to the actual perceptual quantizer Q_p. When the perceptual amplitude non-linearities are taken into account, the corresponding scheme will presumably be more efficient in removing the subjective redundancy from the DFD.

Figure 1 shows the density of quantization levels for each spatial frequency scaled by the total number of quantization levels per frequency, N_f, with and without the amplitude dependencies of the metric (the proposed 2D MPE quantizer and the MPEG-like uniform quantizer). These surfaces show the different distributions of the available quantization levels in the frequency and amplitude plane and represent where the different algorithms concentrate the encoding effort. Figure 2 shows the 2D bit allocation solutions (number of quantization levels per coefficient) using a CSF-based linear metric (MPEG solution) and an amplitude-dependent non-linear metric. Note how the amplitude non-linearities enlarge the bandwidth of the quantizer with regard to the CSF-based case.

3.3 Introducing temporal properties of the HVS in a frame-by-frame motion compensated video coder

The previous perceptual considerations about distances and optimal transform quantizers can be extended to 3D spatio-temporal transforms. The HVS motion perception models extend the 2D spatial filter-bank to non-zero temporal frequencies [45, 46]. The CSF filter is also defined for moving gratings [11], and the contrast discrimination curves for spatio-temporal gratings show the same shape as the curves for still stimuli [12, 34]. By using the 3D CSF and similar non-linear corrections for high amplitudes, the expression of Eq. 4 could be employed to measure differences between local moving patterns. In this way, optimal MPE quantizers could be defined in a spatio-temporal frequency transform domain. However, the frame-by-frame nature of any motion compensated scheme makes the implementation of a 3D transform quantizer in the prediction loop more difficult.

In order to exploit the subjective temporal redundancy removal to some extent, the proposed 2D MPE quantizer can be complemented with a 1D temporal filtering based on the perceptual bit allocation in the temporal dimension. This temporal filter can be implemented by a simple finite impulse response weighting of the incoming error frames. The temporal frequency response of the proposed 1D filter is set proportional to the number of quantization levels that should be allocated in each temporal frequency frame of a 3D MPE optimal quantizer. For each spatio-temporal coefficient, $\mathbf{f} = (f_x, f_t)$, the optimal number of quantization levels is given by Eq. 2. Integrating over the spatial frequency, the number of quantization levels for that temporal frequency is:

$$N_{f_t} = \sum_{f_x} N_{\mathbf{f}} = \frac{1}{2k} \sum_{f_x} \int W_{\mathbf{f}}(a_{\mathbf{f}})^{1/2} \, da_{\mathbf{f}} \tag{5}$$

Figure 3.a shows the number of quantization levels for each spatio-temporal frequency of an optimal 3D non-linear MPE quantizer. Note that the spatial frequency curve for the zero temporal frequency is just the 2D non-linear bit allocation curve of Figure 2. Figure 3.b shows the temporal frequency response which is obtained by integrating over the spatial frequency dimension.

4 Perceptual feed-back in the motion estimation

A characterization of the perceptual quantization process of the encoding domain has an obvious application to the DFD quantizer design, but it may also have interesting effects on the computation of the DVF if the proper feed-back between the motion estimation and the quantization of the DFD in the prediction loop is established.

If all the details of the DFD are considered to be of equal importance, we would have an *unweighted* splitting criterion as in the difference-based criteria [23, 26–29] or as in the spatial entropy-based criterion of Dufaux et al. [21, 22]. However, as the DFD is going to be simplified by some non-trivial quantizer, Q_p, which represents the selective bottleneck of early perception, not every additional detail predicted by a better motion compensation will be significant to the quantizer. In this way, the motion estimation effort has to be focused on the moving regions that contain *perceptually relevant motion information.*

In order to formalize the concept of perceptually relevant motion information, the work of Watson [37] and Daugman [38, 39] on entropy reduction in the HVS should be taken into account. They assume a model of early contrast processing based on a pair (T, Q_p), and suggest that the entropy of the cortical scene representation (a measure of the perceptual entropy of the signal) is just the entropy of the quantized version of the transformed image. Therefore, a simple measure of the perceptual entropy, H_p, of a signal, $\mathbf{A}$, is:

$$H_p(\mathbf{A}) = H\left(Q_p[T(\mathbf{A})]\right) \qquad (6)$$

Using this perceptual entropy measure, an explicit definition of what perceptually relevant motion information is can be proposed. Given a certain motion description, additional motion information, $\Delta H(DVF)$, (more complex quadtree segmentation and more motion vectors) is perceptually relevant only if this additional use of motion information implies a greater reduction in the perceptual entropy of the prediction errors:

$$\Delta H(DVF) < -\Delta H_p(DFD) \qquad (7)$$

i.e., only if the additional motion information adds perceptually significant information to the predictions. Exploiting the fact that the measure of the perceptual entropy of the DFD coincides with the entropy of the output of the perceptually based quantizer, the following *perceptually weighted* entropy-constrained splitting criterion is obtained: *a block of the quadtree structure should be split if,*

$$H(DVF_{split}) + H_p(DFD_{split}) < H(DVF_{nosplit}) + H_p(DFD_{nosplit}) \qquad (8)$$

where $H(DVF)$ is the entropy of the DPCM coded vector field plus the information needed to encode the quadtree structure, and $H_p(DFD)$ is the perceptual entropy of the residual error signal.

Equation 8 has the same form as the criterion proposed by Dufaux and Moscheni [21, 22] except for the way in which the entropy of the DFD is computed. In this case, the unsuitable zero-order entropy of the DFD in the spatial domain is replaced by the entropy measure in the appropriate transform domain [9, 18, 30]. It is interesting to note that the proposed reasoning about the perceptual relevance of the motion information for perceptually matched motion compensation leads to an entropy constrained splitting criterion that takes into account the actual DFD entropy in a natural way.

This perceptual constraint to the motion estimate comes from the particular video coding application in which a human observer assesses the final result. However, the benefits of including the properties of the biological filters in a generic video representation scheme go beyond the optimization of the bit rate for a given subjective distortion.

The particular band-pass shape of human sensitivity [11, 35] gives a scale-dependent measure of the perceptual entropy of the DFD. As some frequency bands (some scales) have more perceptual importance than others, the application of the perceptual criterion results in a different splitting behavior in the different levels of the multigrid structure. Figure 4 qualitatively shows how a band-pass criterion may give a scale-dependent splitting result. In coarse levels of the multigrid (left side figures), the spatial support of the DFD is large due to the displacement of large blocks. The uncertainty relationship that arises between the signal support in frequency and position representations leads to a narrow bandwidth in the case of a large DFD support. Conversely, in the fine levels of the multigrid (right side figures), the DFD is spatially localized giving rise to a broad-band error signal. If the complexity measure is more sensitive to the complexity of the signal in low and medium frequency bands, the splitting criterion will be tolerant in the coarse levels of the multigrid and will be strict in the high resolution levels. This scale-dependent behaviour may be useful for discriminating between significant and noisy motions.

5 Numerical experiments

Several experiments on four standard sequences [47] were carried out using different combinations of quantization schemes and motion estimation algorithms to test the relative contribution of the different proposed alternatives to the final subjective quality at fixed compression ratios.

In the examples below, the benefit of each proposed improvement were separately tested: a comparison of quantizers for a fixed motion estimation algorithm and a comparison of perceptually weighted versus unweighted motion compensations for a fixed quantization were made. The overall performance of the entire encoding schemes proposed here (perceptually weighted variable-size BMA with 2D or 3D MPE redundancy removal) with regard to the techniques used in MPEG-1 and H.263 was also analyzed.

In every experiment, the quantizer was adapted for each group of pictures to achieve the desired bit rate with a fixed maximum perceptual error. Blocks of size 8×8 were used in the fixed-size BMA. A quadtree multigrid scheme with a maximum of five resolution levels (blocks from 64×64 to 4×4) were used in the variable-size BMAs. The n-step displacement search [20, 22] was used in the fixed-size BMA and in every resolution level of the variable-size BMA. The usual correlation was used as a similarity measure. The 1D temporal filter was implemented through a linear-phase FIR filter using least-squares error minimization in the frequency response. A fourth order filter was used to restrict the buffer requirements. Only forward predicted frames and groups of pictures of 10 frames were used, so that the predictive power of the motion field and the ability of the quantizers to avoid error accumulation were emphasized.

Redundancy removal results

The impact of the amplitude non-linearities and the temporal properties in the redundancy reduction process were analyzed comparing the proposed 2D quantizer, with and without temporal filtering, versus the CSF-based MPEG-like quantizer. In the examples presented (Figure 5), the same motion compensation scheme (an unweighted adaptive BMA [21, 22]) was used in every case. Figure 5 shows the reconstructed frame 7 of the RUBIK sequence using different perceptual models to design the quantizer at the same bit rate (200 kbits/sec with QCIF format).

Motion estimation results

The effect of the perceptual feed-back in an entropy-constrained multigrid motion estimation was analyzed comparing the proposed DVF estimation algorithm with the unweighted algorithm of Dufaux et al. which uses the zero-order entropy of the DFD in the spatial domain [21, 22]. Quantitative and qualitative motion estimation results are given. The fixed-resolution BMA has also been included as a useful reference. In every case, the same kind of quantizer (MPEG-like, 2D MPE with linear metric) was used to simplify the DFD.

Table 1 shows the average volume of the flow fields for the algorithms and sequences considered. Figure 6 shows the dependency of the global volume of the encoded signal on the resolution level while the local refinement of the motion estimate is going on. Different starting resolution levels were considered with both splitting criteria.

Figure 7 shows the estimated motion flow and the reconstructed signal (at 200 kbits/sec) for frame 7 of the TAXI sequence for fixed-size BMA and the entropy-constrained variable-size BMA with different (unweighted and perceptually weighted entropy measures. Figures 7.d to 7.f display a representation of the flow results that highlight its meaningfulness. Each block of the final quadtree structure was classified as one of the still or moving objects according to its nearest displacement neighbor in order to obtain this representation.

Overall results

Three basic encoding configurations were compared to test the joint performance of the different algorithms considered:

- MPEG1-like scheme with no balance between DFD and DVF: fixed-size BMA and a spatial uniform CSF-based quantizer which restricts the MPE according to a linear metric.
- H.263-like scheme with an unweighted entropy-constrained motion estimation and a uniform CSF-based quantizer.
- The proposed schemes (with 2D or 3D subjective redundancy removal) which use a perceptually weighted entropy-constrained motion estimation and a spatial non-uniform quantizer which restricts the MPE according to a non-linear metric. The 3D variant of this basic scheme includes some perceptual temporal features through the heuristic filter of Figure 3.b.

Some examples of the results are shown in Figures 9 and 10 which display the reconstruction of frame 7 of the RUBIK sequence and frame 7 of the TAXI sequence for the different compression schemes at 200 kbits/sec (QCIF format).

6 Discussion

6.1 Effect of amplitude non-linearities on quantizer results

As expected [15], the quantizers that include the amplitude non-linearities outperform the behaviour of the CSF-based MPEG-like quantizer. If the DFD is simplified according to a more accurate perception model, the perceptually relevant information is better described, so the quantization errors are less visible in the reconstructed frames. The subjective quality of the reconstructed sequence drops faster in the threshold-based MPEG-like quantizer case in such a way that more intracoded frames would be necessary to keep the same quality with the consequent increase in the bit rate.

The introduction of the low-pass perceptual temporal filter smooths the reconstructed sequence reducing the blocking effect and some flickering of the 2D approach, giving a more pleasant overall quality. However, despite the eventual subjective gain due to the temporal filtering, the example in Figure 5 shows that the key factor that improves the subjective appearance is the introduction of the 2D perceptual amplitude non-linearities in the quantizer. The spreading of the effective bandwidth of the non-linear quantizer (see Figure 2) keeps some high frequency details of the DFD. This avoids the rapid degradation of the reconstructed sequence. Note that, despite the bits allocated in the high frequency regions, not every high frequency contribution will be preserved because it will also depend on its amplitude.

6.2 *Effect of perceptual feed-back on motion representation*

The main (qualitative) aim of locally adaptive motion estimation techniques in video coding is to obtain a motion description which has greater predictive power, i.e. a motion description which results in better temporal redundancy removal. The particular (quantitative) aim of the entropy-constrained control of the adaptive motion estimation is to find an optimal bit allocation between the DVF and the DFD.

Therefore, there are two possible approaches to the analysis of the results of the entropy-constrained motion estimation algorithms: in terms of bit rate [9, 18, 30, 31] and in terms of the usefulness of the motion representation.

As previously reported from the quantitative (bit rate-based) point of view [9, 18, 30], the spatial entropy algorithm is not optimal because, by definition, it does not use the actual volume of the encoded signal, but rather an estimate that may be inaccurate if the DFD encoding process is non-trivial. In contrast, it can be seen (Figure 6) that the proposed criterion does minimize the volume of the signal (in the meaningful encoding domain) no matter what the starting resolution level is. With the proposed criterion, the motion information needed for compensation is reduced by factors of 2 and 10 with regard to spatial-entropy based variable-size BMA and fixed-size BMA, respectively (Table. 1).

The important question is whether or not the quantitative savings in the DVF volume imply a qualitative relevant benefit. There are two possibilities for this to be true. One possibility is that the savings in the motion description implies a better DFD encoding and better subjective quality at the same compression ratio, which is the rationale behind the bit rate-based analysis. The other, more interesting possibility is that the introduction of the perceptual feed-back in the motion estimation induces desirable qualitative features in the final flow.

The analysis of the reconstructed signal gives us information about the effectiveness of the savings in the motion description from the *quantitative* point of view. The properties of the final flows to represent the motion in the scene give us information about the benefits of a perceptually weighted criterion from the *qualitative* point of view.

It can be seen that, regardless of the splitting criterion, the adaptive schemes give a more predictive motion estimate that implies a better reconstruction at a given bit rate (compare Figure 7.a with Figures. 7.b and 7.c). It is also worth noting that in the adaptive scheme with the proposed criterion, less motion vectors are needed to describe the motion in the scene obtaining the same quality in the reconstructed frame. From the *quantitative* point of view, the significant change in the case of the TAXI sequence is the reduction of the DVF volume from 13% in the case of the fixed-size BMA to less than 2% in the adaptive schemes. In the adaptive cases, the (same) quantizer can take advantage of these savings (of 10%) to give a better reconstruction. In contrast, the additional benefit of the proposed algorithm (here, an additional 1%) is too small to be useful to the quantizer. This behaviour was the same in all the explored cases, where the average additional benefit of the proposed motion estimation was around 2 or 3%. This result suggests that the advantages of an optimal algorithm (in the sense of measuring the entropy in the appropriate domain) gives rise to a lower complexity motion description, but this does not make a significant improvement in the reconstructed signal.

The actual advantages of the proposed motion estimation algorithm are not related to its quantitative optimality but to the qualitative effects on the final flow.

In general, with the proposed splitting algorithm, less false alarms are obtained and significative motion is better described. Note how in the static area at the bottom of the TAXI frame (Figures 7.d and 7.e), the spatial-entropy-based algorithm gives the same false alarm as uniform resolution BMA due to an unadequate increase of the resolution while the motion of the incoming vehicle on the right is not properly detected.

As stated before, the effect of the perceptual feed-back is a more conservative splitting strategy which gives rise to more robust estimates. This is because perceptual quantization shows a band-pass behavior (Figures 2 and 4) neglecting certain texture components of the prediction error that would have been considered by the unweighted spatial entropy criterion. In fact, resolution is increased only if the corresponding moving structures are perceptually relevant. With the unweighted criterion, too many blocks are unnecesarily split. This accounts for the increase in the entropy of the signal after a local mimimum at medium resolution (Figure 6).

The benefits of this splitting strategy are also reflected in the coherence of the quadtree segmentated representation (Figures 7.d to 7.f). The histogram of the estimated displacements can account for these results (Figure 8). While the fixed size BMA estimates are noisy (the distribution of the cluster corresponding to each moving object has a large variance), the scale-dependent behavior of the proposed splitting criterion gives rise to robust motion estimates that concentrate around the actual displacements, so it obtains the sharpest peaks. It is clear that this behavior may simplify the task of a hypothetical higher level scene interpretation based on accurate speed measurements or motion-based segmentations. This improved motion representation may be interesting as a first stage in a model-based video coding framework.

6.3 Overall results

The entire set of experiments with the different global encoding configurations using different sequences shows a clear improvement in the subjective quality obtained with the proposed 2D or 3D schemes with regard to the MPEG-1 or the (unweighted) H.263 schemes (Figures 9 and 10).

The relative contribution of the motion estimation and DFD quantization improvements can be evaluated by comparing their joint effect on the same sequences with regard to the results discussed in subsections 6.1 and 6.2.

The overall behaviour is consistent with the separate results discussed above: while the improved motion representation due to the proposed splitting strategy reduces the DVF volume to a certain extent, the main differences in the subjective quality are due to the use of a better quantizer including a more accurate description of the HVS redundancy removal. The proposed 2D quantizer reduces the annoying busy artifacts around the moving objects and the temporal filtering reduces the visibility of temporal impulsional artifacts (in individual frames) at the cost of a slight smoothing of the small moving edges.

7 Final remarks

A multigrid motion compensated video coding scheme based on the current models of HVS contrast discrimination has been presented. These models account for the non-uniform nature of the HVS redundancy removal in the spatio-temporal frequency domain. The basic idea is to design the entire encoding process (motion estimation and prediction error simplification) in order to preserve no more than the subjectively relevant information at a given subjective distortion level.

This aim leads to the design of a perceptually inspired transform quantizer in a natural way. Moreover, a perceptually weighted entropy-constrained criterion to refine the quadtree structure for motion estimation is obtained by using the perceptual quantizer to control the adaptive flow estimation. In this way, the excess-effort in the motion description (predicting details that are going to be discarded by the quantizer) is avoided, and a perceptual feed-back is introduced in the motion estimation.

Using an explicit design criterion in the transform (DCT) domain –restricting the *maximum perceptual error*– and an appropriate perceptual distortion metric, a set of metric-dependent expressions have been proposed to implement the quantizer that simulates the HVS redundancy removal processes. These expressions lead to uniform MPEG-like quantizers if a simple (linear CSF-based) metric is used. In the proposed implementation, a more general metric which includes the perceptual amplitude non-linearities and some temporal properties has been used.

The results show that, on one hand, the proposed non-uniform quantizer scheme leads to a better subjective quality in the reconstructed signal than the CSF-based MPEG-like quantizers at the same bit rates. On the other hand, the perceptual feed-back leads to a scale-dependent motion refinement strategy that gives more robust motion estimates than an unweighted entropy-based splitting criterion. It has been shown that the proposed motion estimation algorithm may facilitate segmentation tasks and higher level scene interpretation that are of interest in model-based video coding. The comparison of the proposed scheme versus a H.263 scheme with unweighted motion refinement and MPEG-like quantization shows the overall subjective superiority of the proposed improvements.

References

[1] D. LeGall. MPEG: A video compression standard for multimedia applications. *Communications of the ACM*, 34(4):47–58, 1991.

[2] G. Tziritas and C. Labit. *Motion Analysis for Image Sequence Coding*. Elsevier Science, Amsterdam, 1994.

[3] A.M. Tekalp. *Digital Video Processing*. Prentice Hall, Upper Saddle River, NJ, 1995.

[4] ITU-Telecommunication Standarization Sector. Draft recommendation H.263, 1994.

[5] ISO/IEC JTC1/SC29/WG11 N1909. Overview of the MPEG-4 version 1 standard, 1997.

[6] M. Bierling. Displacement estimation by hierarchical block-matching. *Proceedings of the SPIE, Conf. Visual Communications and Image Processing*, 1001:942–951, 1988.

[7] J. Malo, F. Ferri, J. Albert, J.Soret, and J.M. Artigas. The role of perceptual contrast non-linearities in image transform coding. *Image & Vision Computing*, Accepted April 1999.

[8] J. Malo, F. Ferri, J. Albert, and J. Soret. Comparison of perceptually uniform quantization with average error minimization in image transform coding. *Electronics Letters*, 35(13):1067–1068, 1999.

[9] J. Malo, F. Ferri, J. Albert, and J.M. Artigas. Splitting criterion for hierarchical motion estimation based on perceptual coding. *Electronics Letters*, 34(6):541–543, 1998.

[10] G.E Legge. A power law for contrast discrimination. *Vision Research*, 18:68–91, 1981.

[11] D.H. Kelly. Motion and vision II: Stabilized spatiotemporal threshold surface. *Journal of the Optical Society of America*, 69(10):1340–1349, 1979.

[12] B.L. Beard, S. Klein, and T. Carney. Motion thresholds can be predicted from contrast discrimination. *Journal of the Optical Society of America A*, 14(9):2449–2470, 1997.

[13] B. Girod. Motion compensation: Visual aspects, accuracy and fundamental limits. In M.I. Sezan and R.L. Lagendijk, editors, *Motion Analysis and Image Sequence Processing*, 1993.

[14] N. Jayant, J. Johonston, and R. Safranek. Signal compression based on models of human perception. *Proceedings IEEE*, 81(10):1385–1422, 1993.

[15] J. Malo, A.M. Pons, and J.M. Artigas. Bit allocation algorithm for codebook design in vector quantization fully based on human visual system non-linearities for suprathreshold contrasts. *Electronics Letters*, 31(15):1229–1231, 1995.

[16] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Press, Boston, 1992.

[17] D.W. Lin, M.H. Wang, and J.J. Chen. Optimal delayed coding of video sequences subject to a buffer size constraint. *Proceedings of the SPIE, Conf. Visual Communications and Image Processing*, 2094(0):223–234, 1993.

[18] G.M. Schuster and A.K. Katsaggelos. *Rate-Distortion Based Video Compression.* Kluwer Academic Publishers, Boston, 1997.

[19] B. Macq. Weighted optimum bit allocations to orthogonal transforms for picture coding. *IEEE Journal on Selected Areas in Communications*, 10(5):875–883, 1992.

[20] H.G. Musmann, P. Pirsch, and H.J. Grallert. Advances in picture coding. *Proceedings IEEE*, 73(4):523–548, 1985.

[21] F. Moscheni, F. Dufaux, and H. Nicolas. Entropy criterion for optimal bit allocation between motion and prediction error information. *Proceedings of the SPIE, Conf. Visual Communications and Image Processing*, 2094:235–242, 1993.

[22] F. Dufaux, F. Moscheni, and M. Kunt. Motion estimation techniques for digital TV: A review and new contribution. *Proceedings IEEE*, 83(6):858–876, 1995.

[23] M.H. Chan, Y.B. Yu, and A.G. Constantinides. Variable size bock matching motion compensation with applications to video coding. *Proceedings IEE, Vision Image and Signal Processing*, 137(4):205–212, 1990.

[24] E. Reusens et al. Dynamic approach to visual data compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(1):197–211, 1997.

[25] F. Dufaux and F. Moscheni. Segmentation-based motion estimation for second generation video coding techniques. In L. Torres and M. Kunt, editors, *Video Coding: A Second Generation Approach*, 1996.

[26] F. Dufaux, I. Moccagatta, B. Rouchouze, T. Ebrahimi, and M. Kunt. Motion-compensated generic coding of video based on multiresolution data structure. *Optical Engineering*, 32(7):1559–1570, 1993.

[27] J. Li and X. Lin. Sequantial image coding based on multiresolution tree architecture. *Electronics Letters*, 29(17):1545–1547, 1993.

[28] V. Seferidis and M. Ghanbari. Generalised block-matching motion estimation using quad-tree structured spatial decomposition. *Proceedings IEE, Vision Image and Signal Processing*, 141(6):446–452, 1994.

[29] M.H. Lee and G. Crebbin. Image sequence coding using quadtree-based block-matching motion estimation and classified vector quantization. *Proceedings IEE, Vision Image and Signal Processing*, 141(6):453–460, 1994.

[30] G.M. Schuster and A.K. Katsaggelos. A video compression scheme with optimal bit allocation among segmentation, motion and residual error. *IEEE Transactions on Image Processing*, 6(11):1487–1502, 1997.

[31] J. Lee. Joint optimization of block size and quantization for quadtree-based motion estimation. *IEEE Transactions on Image Processing*, 7(6):909–912, 1998.

[32] A.B. Watson and J.A. Solomon. A model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A*, 14:2379–2391, 1997.

[33] H.R. Wilson. Pattern discrimination, visual filters and spatial sampling irregularities. In M.S. Landy and J.A. Movshon, editors, *Computational Models of Visual Processing*, pages 153–168, Massachussets, 1991. MIT Press.

[34] E. Martinez-Uriegas. Color detection and color contrast discrimination thresholds. In *Proceedings of the OSA Annual Meeting ILS–XIII*, page 81, Los Angeles, 1997.

[35] J. Malo, A.M. Pons, and J.M. Artigas. Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain. *Image & Vision Computing*, 15(7):535–548, 1997.

[36] A.M. Pons, J. Malo, J.M. Artigas, and P. Capilla. Image quality metric based on multidimensional contrast perception models. *Displays*, Accepted Feb. 1999.

[37] A.B. Watson. Efficiency of a model human image code. *Journal of Optical Society of America A*, 4(12):2401–2417, 1987.

[38] J.G. Daugman. Complete discrete 2D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1169–1179, 1988.

[39] J.G. Daugman. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Transactions on Biomedical Engineering*, 36:107–114, 1989.

[40] A.J. Ahumada and H.A. Peterson. Luminance-model-based DCT quantization for color image compression. volume 1666 of *Proceedings of the SPIE*, pages 365–374, 1992.

[41] J.A. Solomon, A.B. Watson, and A.J. Ahumada. Visibility of DCT basis functions: effects of contrast masking. In *Proceedings of Data Compression Conference, Snowbird, Utah*, IEEE Computer Society Press, pages 361–370, 1994.

[42] A.B. Watson. DCT quantization matrices visually optimized for individual images. In B.E. Rogowitz, editor, *Human Vision, Visual Processing and Digital Display IV*, volume 1913, 1993.

[43] S.P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):127–135, 1982.

[44] N.B. Nill and B.R. Bouzas. Objective image quality measure derived from digital image power spectra. *Optical Engineering*, 32(4):813–825, 1992.

[45] D.J Heeger. Model for the extraction of image flow. *Journal of the Optical Society of America A*, 4:1455–1471, 1987.

[46] A.B. Watson and A.J. Ahumada. Model of human visual motion sensing. *Journal of the Optical Society of America A*, 2:322–342, 1985.

[47] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

Table 1
Relative volume of the motion flow (in %) with regard to the total volume of the encoded signal (200 kbits/sec)

	TAXI	RUBIK	YOSEMITE	TREES	Average
Fixed-size BMA	12.57	21.71	46.19	39.21	30 ± 8
Unweighted entropy	1.89	3.29	8.69	4.97	4.7 ± 1.5
Perceptual entropy	0.94	1.87	4.32	3.24	2.6 ± 0.7

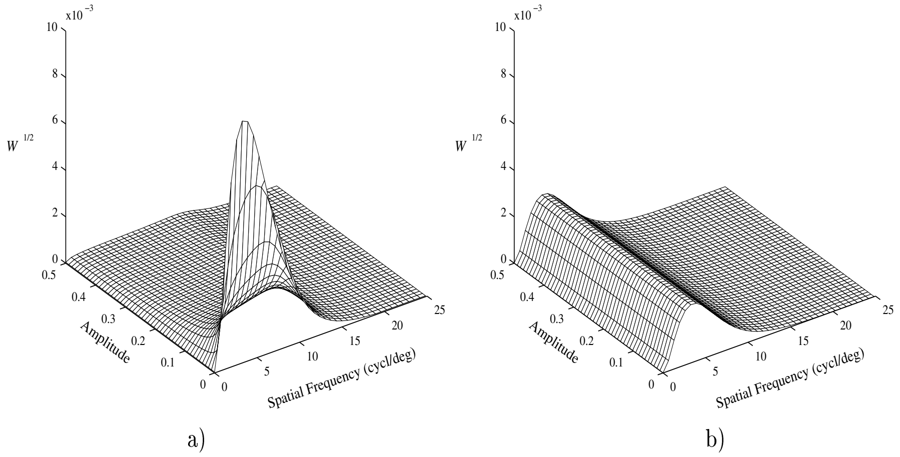


Fig. 1. *Density of quantization levels in the frequency and amplitude plane for a) non-linear MPE and, b) linear MPE quantizers. Note that in the MPE approach the final density is proportional to the metric of the domain:* $N_f \cdot \lambda_f(a_f) \propto W(a_f)^{1/2}$.

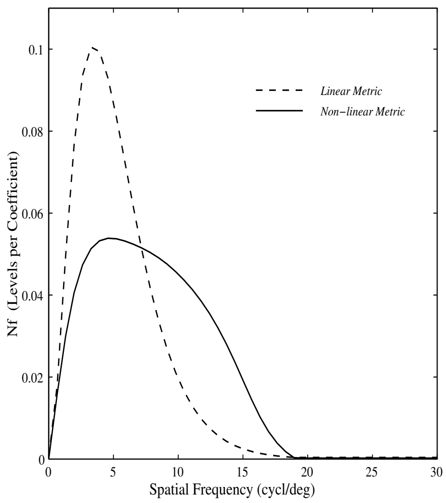
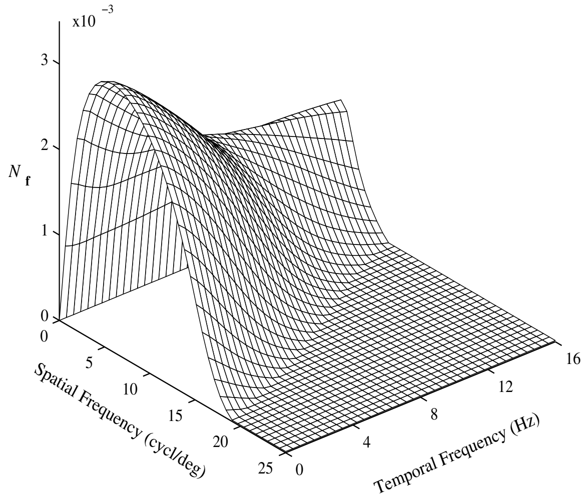
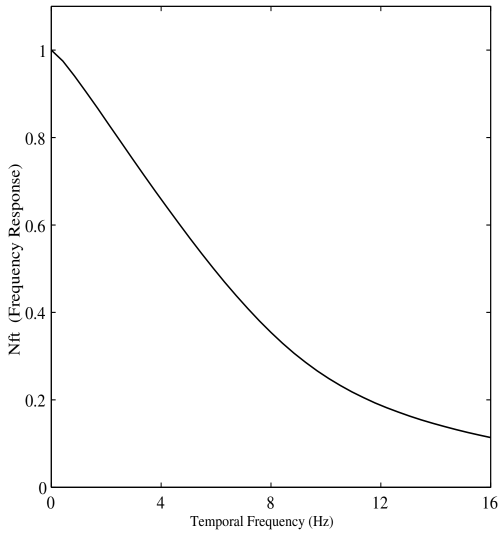


Fig. 2. *Bit allocation results (Relative number of quantization levels per coefficient) for the linear MPE (MPEG-like case), and for the non-linear MPE case.*



a)



b)

Fig. 3. *a) Non-linear MPE bit allocation results in the 3D spatio-temporal frequency domain (Relative number of quantization levels per coefficient). b) Frequency response of the perceptual temporal filter, proportional to* N_{f_t}.

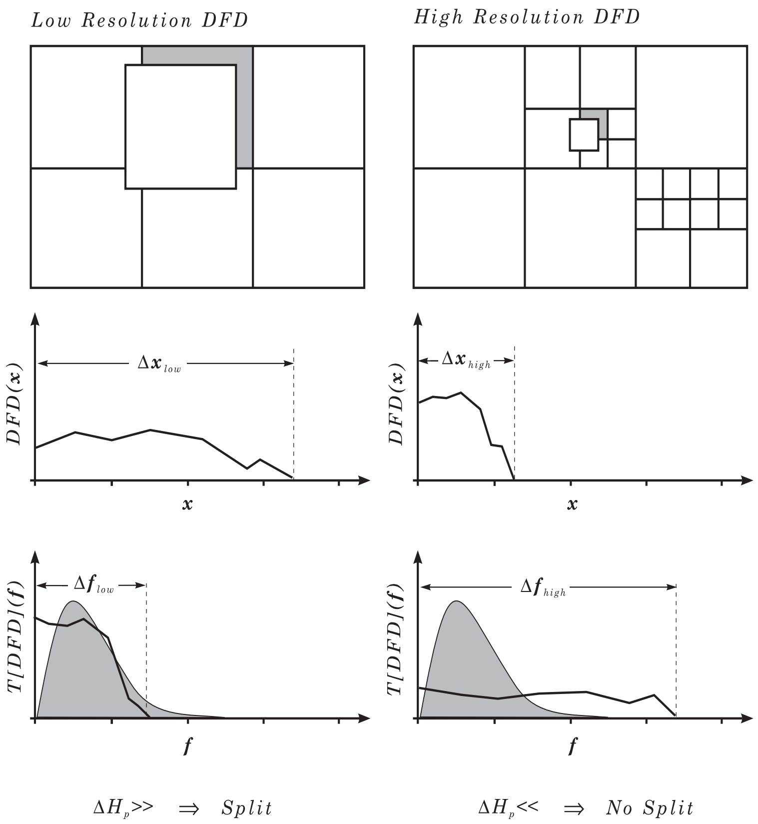
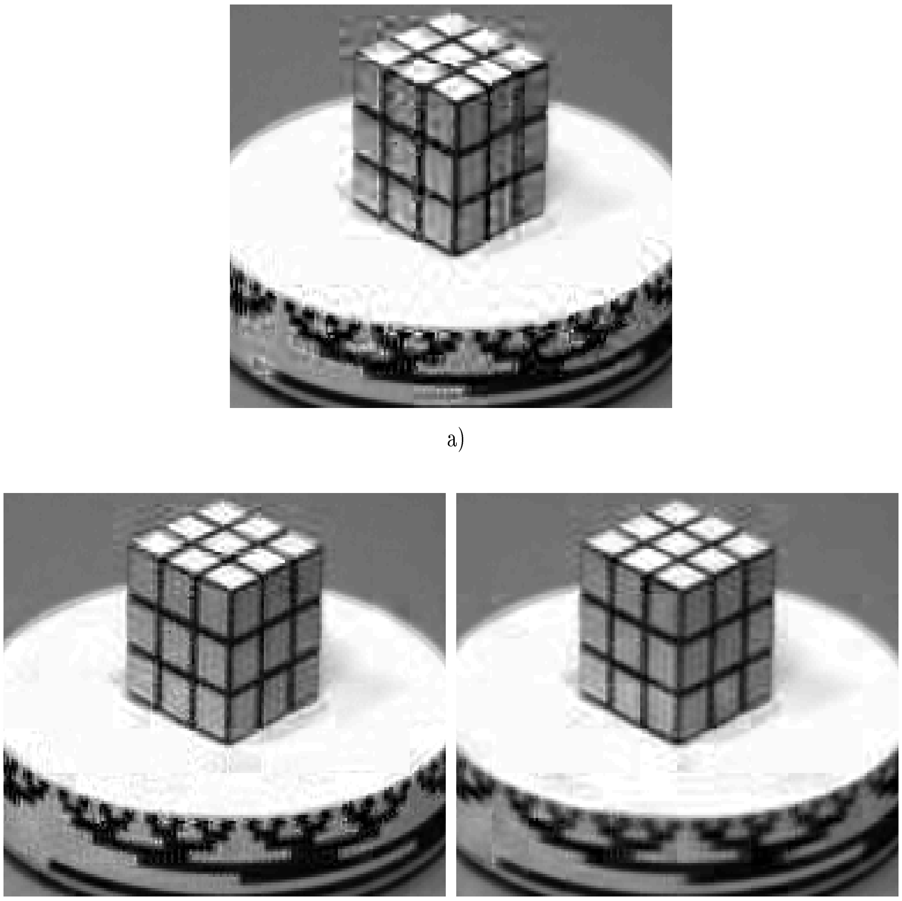


Fig. 4. *Scale-dependent splitting strategy due to perceptual feed-back. For a given energy and resolution level, the spatial extent and the frequency bandwidth of the DFD are related by the uncertainty relation, $\Delta x \cdot \Delta f = k$. The bandwidth of the DFD will depend on the resolution, giving rise to a different splitting behaviour when using a band-pass splitting criterion.*



a)

b) c)

Fig. 5. *Quantization results with a fixed motion estimation algorithm (unweighted variable-size BMA). a) 2D Linear MPE, uniform MPEG-like quantization, b) 2D Non-linear MPE, c) 2D Non-linear MPE and temporal filtering.*

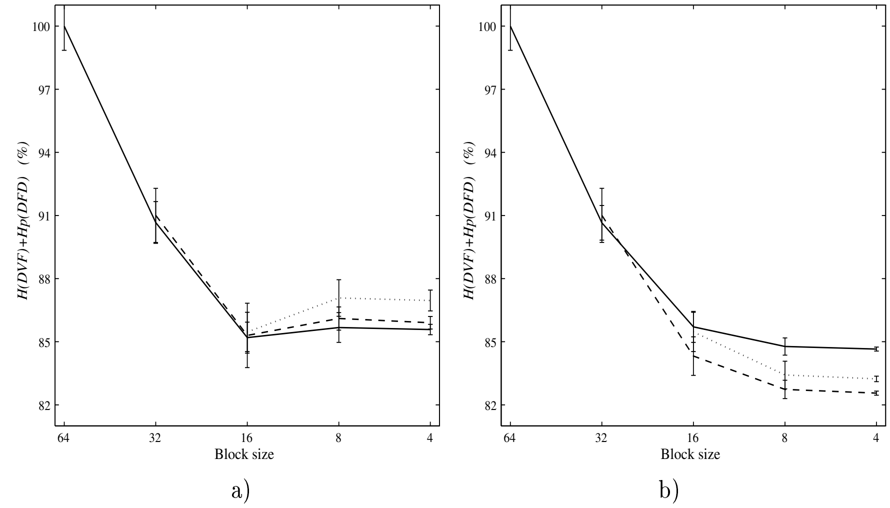


Fig. 6. *Volume of flow and prediction errors while refining the motion estimate. The values are given as a percentage of the total entropy at the lowest resolution level. a) Unweighted spatial entropy splitting criterion. b) Perceptual entropy splitting criterion. Note that the proposed criterion behaves monotonically and achieves the lowest values (for a fixed MPE) regardless of the starting level.*

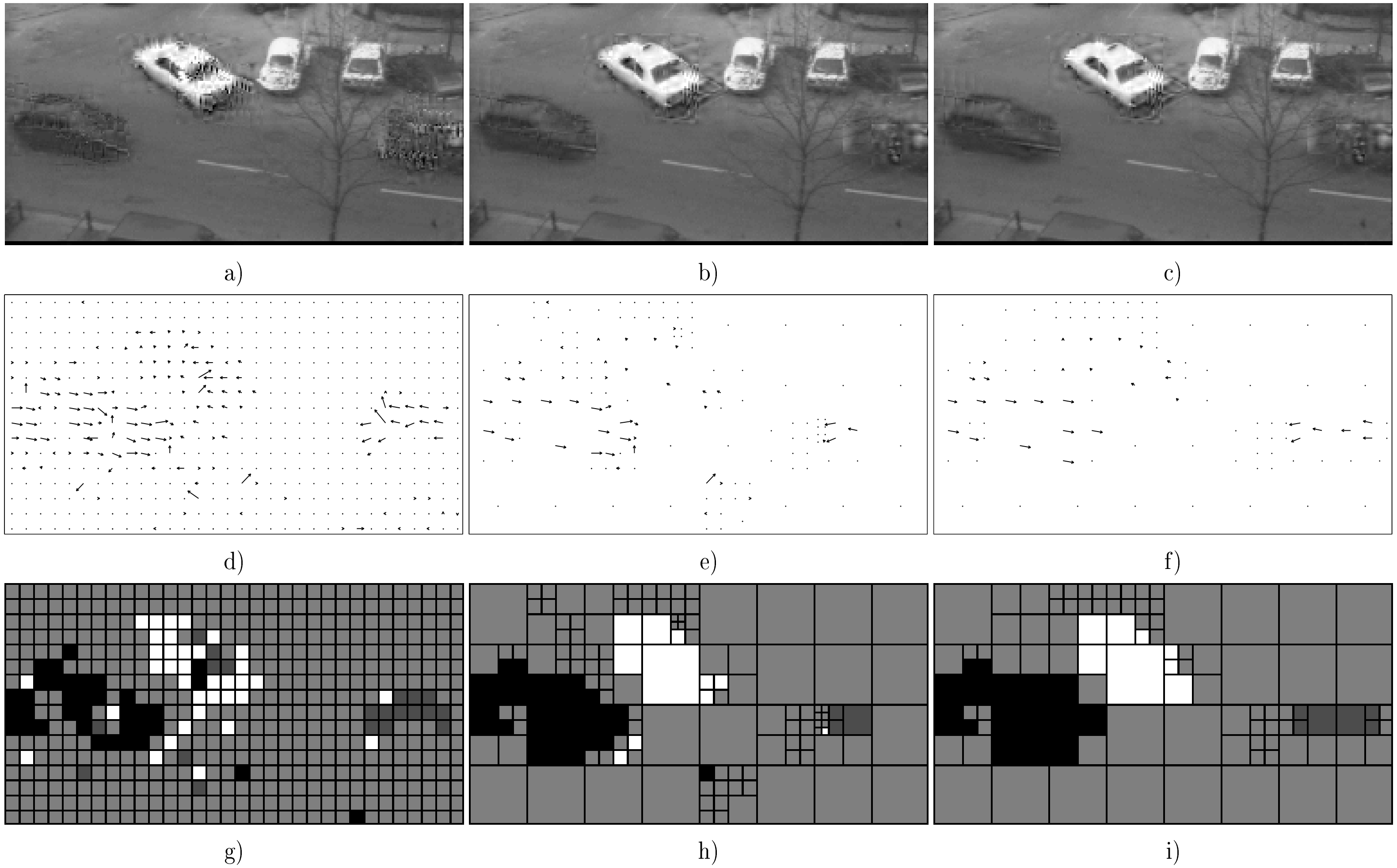


Fig. 7. Motion estimation results for a fixed uniform MPEG-like quantization scheme. Reconstructed signals (a-c), motion flows (d-f), and quadtree segmentated representation (g-i).

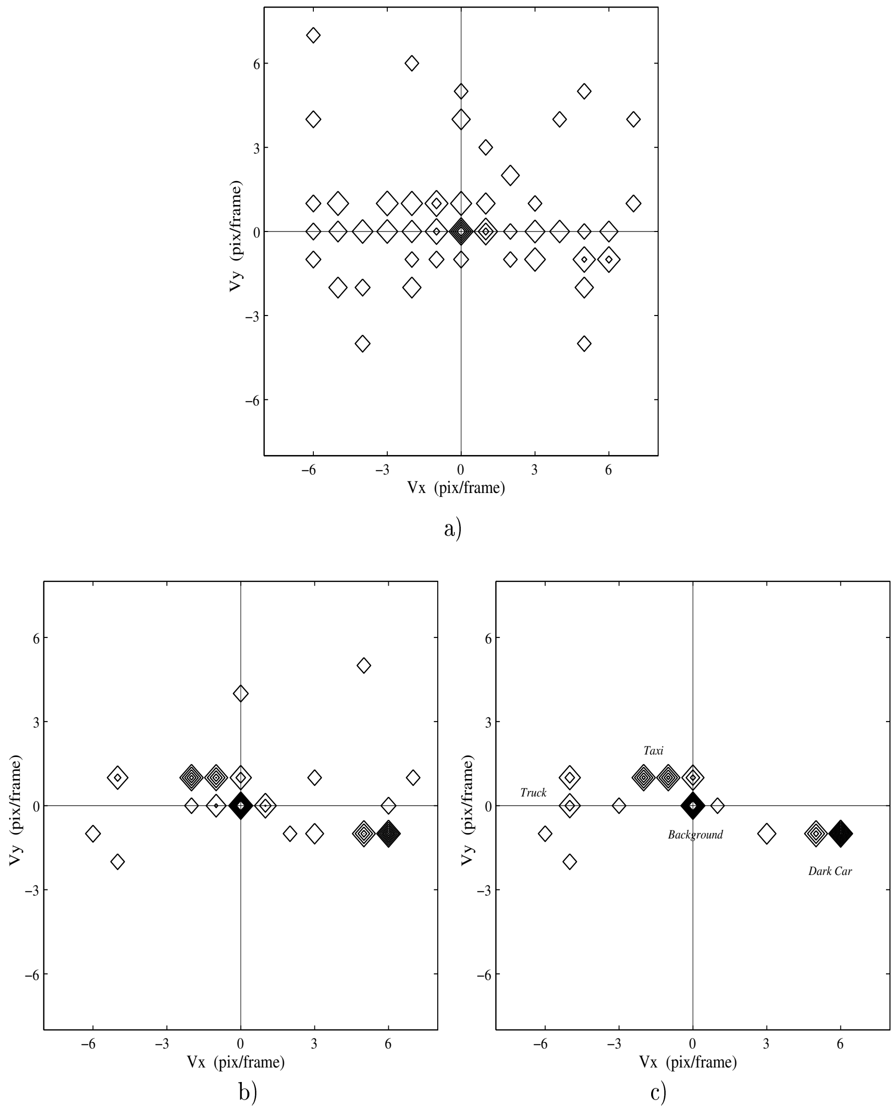
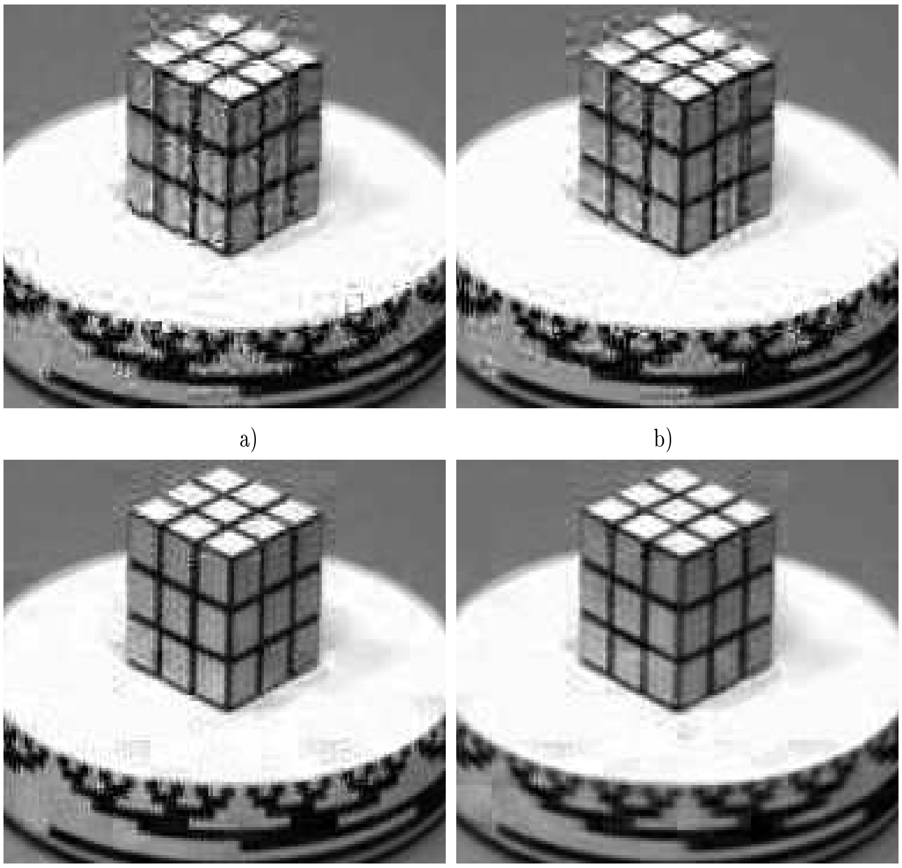


Fig. 8. *Contour view of the displacement histograms obtained by the different motion estimation algorithms. a) Fixed-size BMA, b) Unweighted adaptive BMA, c) Perceptually weighted adaptive BMA.*



a) b)

c) d)

Fig. 9. *Overall reconstruction results with the RUBIK sequence (200 kbits/sec) using previously reported encoding configurations (a-b) and the proposed 2D or 3D alternatives (c-d). a) Fixed size BMA for motion estimation and MPEG-like quantization (linear MPE). b) Unweighted variable-size BMA and MPEG-like quantization (linear MPE). c) Perceptually weighted variable-size BMA and 2D non-linear MPE quantization. d) Perceptually weighted variable-size BMA and 2D non-linear MPE quantization and temporal filtering.*

Fig. 10. *Zoom of the overall reconstruction results with the TAXI sequence (200 kbits/sec) using previously reported encoding configurations (a-b) and the proposed 2D or 3D alternatives (c-d). a) Fixed size BMA for motion estimation and MPEG-like quantization (linear MPE). b) Unweighted variable-size BMA and MPEG-like quantization (linear MPE). c) Perceptually weighted variable-size BMA and 2D non-linear MPE quantization. d) Perceptually weighted variable-size BMA and 2D non-linear MPE quantization and temporal filtering. The smoothing of the small moving objects in the 3D approach is not as noticeable in the actual (moving) sequences.*

Bibliografía

[1] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H. Freeman and Co., New York, 1978.

[2] B.A. Wandell. *Foundations of Vision.* Sinauer Assoc. Publish., Massachusetts, 1995.

[3] A.N. Akansu and R.A. Haddad. *Multiresolution Signal Decomposition.* Academic Press, Boston, 1992.

[4] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis and Machine Vision.* Intl. Thomson Computer Press, London, 1999.

[5] A. Rosenfeld. Computer vision: A source of models for biological visual process? *IEEE Transactions on Biomedical Engineering*, 36:93–96, 1989.

[6] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, 1973.

[7] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, Boston, 1990.

[8] R.J. Clarke. *Transform Coding of Images.* Academic Press, New York, 1985.

[9] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Press, Boston, 1992.

[10] A.M. Tekalp. *Digital Video Processing.* Prentice Hall, Upper Saddle River, NJ, 1995.

[11] G. Tziritas and C. Labit. *Motion Analysis for Image Sequence Coding.* Elsevier Science, Amsterdam, 1994.

[12] G.M. Schuster and A.K. Katsaggelos. *Rate-Distortion Based Video Compression.* Kluwer Academic Publishers, Boston, 1997.

[13] J.G. Daugman. Six formal properties of two-dimensional anisotropic visual filters: Structural principles and frequency/orientation selectivity. *IEEE Transactions on Systems, Man and Cybernetics*, 13(5):882–887, 1986.

[14] J.G. Daugman. Complete discrete 2D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1169–1179, 1988.

[15] J.G. Daugman. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Transactions on Biomedical Engineering*, 36:107–114, 1989.

[16] D. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

[17] A.B. Watson. Detection and recognition of simple spatial forms. In O.J. Braddick and A.C. Sleigh, editors, *Physical and Biological Processing of Images*, volume 11 of *Springer Series on Information Sciences*, pages 100–114, Berlin, 1983. Springer Verlag.

[18] A.B. Watson and A.J. Ahumada. A hexagonal orthogonal oriented pyramid as a model of image representation in visual cortex. *IEEE Transactions on Biomedical Engineering*, 36:97–106, 1989.

[19] H.R. Wilson. Pattern discrimination, visual filters and spatial sampling irregularities. In M.S. Landy and J.A. Movshon, editors, *Computational Models of Visual Processing*, pages 153–168, Massachussets, 1991. MIT Press.

[20] M.S. Landy and J.R. Bergen. Texture segregation and orientation gradient. *Vision Research*, 31(4):679–691, 1991.

[21] M. Clark and A.C. Bovik. Experiments in segmenting texton patterns using localized spatial filters. *Pattern Recognition*, 22:707–717, 1989.

[22] A.C. Bovik, M. Clark, and W.S. Geisler. Multichannel texture analysis using localized spatial textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):55–73, 1990.

[23] M. Porat and Y.Y. Zeevi. The generalized Gabor scheme for image representation in bilogical and machine vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:452–468, 1988.

[24] M. Porat and Y.Y. Zeevi. Localized texture processing in vision: Analysis and synthesis in the gaborian space. *IEEE Transactions on Biomedical Engineering*, 36:115–129, 1989.

[25] J. Portilla, R. Navarro, O. Nestares, and A. Tabernero. Texture synthesis-by-analysis based on a multi-scale early-vision model. *Optical Engineering*, 35:2403–2417, 1996.

[26] R. Navarro, A. Tabernero, and G. Cristobal. Image representation with Gabor wavelets and its applications. *Advanced in Imaging and Electron Physics*, 97:1–84, 1996.

[27] T.V. Papathomas, R.S. Kashi, and A. Gorea. A human vision based computational model for chromatic texture segregation. *IEEE Transactions on Systems, Man & Cybernetics, B*, 27(3):428–440, 1997.

[28] A.K. Jain, N.K. Ratha, and S. Lakshmanan. Object detection using Gabor filters. *Pattern Recognition*, 30(2):295–309, 1997.

[29] E.H. Adelson and J.R. Bergen. Spatio-temporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284–299, 1985.

[30] A.B. Watson and A.J. Ahumada. Model of human visual motion sensing. *Journal of the Optical Society of America A*, 2:322–342, 1985.

[31] D.J Heeger. Model for the extraction of image flow. *Journal of the Optical Society of America A*, 4:1455–1471, 1987.

[32] D.J Heeger. Optical flow using spatiotemporal filters. *International Journal of Computer Vision*, 1:279–302, 1987.

[33] T.B. Lawton. Outputs of paired Gabor filters summed across the background frame of reference predict the direction of movement. *IEEE Transactions on Biomedical Engineering*, 36(1):130–139, 1989.

[34] J. Lubin. *The Use of Psychophysical Data and Models in the Analysis of Display System Performance.* In A.B. Watson, editor, *Digital Images and Human Vision*, pages 163–178, Massachusetts, 1993. MIT Press.

[35] S. Daly. Visible differences predictor: An algorithm for the assessment of image fidelity. In A.B. Watson, editor, *Digital Images and Human Vision*, pages 179–206, Massachusetts, 1993. MIT Press.

[36] A.J. Ahumada. Computational image quality metrics: A review. In J. Morreale, editor, *Intl. Symp. Dig. of Tech. Papers, Sta. Ana CA*, volume 25 of *Proceedings of the SID*, pages 305–308, 1993.

[37] P.C. Teo and D.J. Heeger. Perceptual image distortion. *Proceedings of the SPIE*, 2179:127–139, 1994.

[38] H.R. Wilson, D. Levi, L. Maffei, J. Rovamo, and R. DeValois. The perception of form. In L. Spillmann and J.S Werner, editors, *Visual Perception: The Neurophysiological Foundations*, pages 231–272, San Diego, 1990. Academic Press.

[39] A.B. Watson and J.A. Solomon. A model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A*, 14:2379–2391, 1997.

[40] P.J. Burt and E.J Adelson. The laplacian pyramid as a compact image code. *IEEE Transaction on Communications*, 31:532–540, 1983.

[41] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[42] J.W. Woods and S.D. O'Neil. Subband coding of images. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(5):1278–1288, 1986.

[43] H.S. Malvar and D.H. Staelin. The LOT: Transform coding without blocking effects. *IEEE Transactions on Acoustics Speech and Signal Processing*, 37(4):553–559, 1989.

[44] J.G. Daugman. Two-dimensional spectral analysis of cortical recptive field profiles. *Vision Research*, 20:847–856, 1980.

[45] J.G. Daugman. Spatial visual channels in the fourier plane. *Vision Research*, 24(9):891–910, 1984.

[46] S. Marcelja. Mathematical description of the response of simple cortical cells. *Journal of the Optical Society of America*, 70(11):1297–1300, 1980.

[47] A.B. Watson. The cortex transform: Rapid computation of simulated neural images. *Computer Vision, Graphics and Image Processing*, 39:311–327, 1987.

[48] D.E Pearson. Developments in model-based video coding. *Proceedings IEEE*, 83(6):892–906, 1995.

[49] H.G. Musmann, M. Hotter, and J. Ostermann. Object-oriented analysis synthesis coding of moving images. *Signal Processing: Image Communication*, 1(1):117–138, 1989.

[50] R. Forchheimer and T. Kronander. Image coding: from waveforms to animation. *IEEE Transactions on Acoustics Speech and Signal Processing*, 37:2008–2023, 1989.

[51] M. Kunt, A. Ikonomopoulos, and M. Kocher. Second-generation image coding techniques. *Proceedings IEEE*, 73:549–574, 1985.

[52] N. Jayant, J. Johonston, and R. Safranek. Signal compression based on models of human perception. *Proceedings IEEE*, 81(10):1385–1422, 1993.

[53] A.B. Watson. Perceptual-components architecture for digital video. *Journal of the Optical Society of America A*, 7(10):1943–1954, 1990.

[54] M.D Fairchild. *Color Appearance Models*. Addison-Wesley, New York, 1997.

[55] N. Abramson. *Information Theory and Coding*. McGraw-Hill, New York, 1964.

[56] IEEE Signal Processing Magazine. Special issue on rate-distortion methods for image and video compression, November 1998.

[57] T. Berger. *Rate Distortion Theory.* Prentice-Hall, Englewood Cliffs, NJ., 1971.

[58] H.G. Musmann, P. Pirsch, and H.J. Grallert. Advances in picture coding. *Proceedings IEEE*, 73(4):523–548, 1985.

[59] F. Dufaux, F. Moscheni, and M. Kunt. Motion estimation techniques for digital TV: A review and new contribution. *Proceedings IEEE*, 83(6):858–876, 1995.

[60] J. Lee. Joint optimization of block size and quantization for quadtree-based motion estimation. *IEEE Transactions on Image Processing*, 7(6):909–912, 1998.

[61] C. Hwang, S. Venkatraman, and K.R. Rao. Human visual system weighted progressive image transmission using lapped orthogonal transform classified vector quantization. *Optical Engineering*, 32(7):1525–1530, 1993.

[62] IEEE Transactions on Information Theory. Special issue on wavelet transforms and multiresolution signal analysis, March 1992.

[63] R.R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, 1992.

[64] H.S. Wu, R.A. King, and R.I. Kitney. Improving the performance of the quadtree-based image aproximation via the generalized DCT. *Electronics Letters*, 29(10):887–888, 1993.

[65] C. Herley, Z. Xiong, K. Ramchandran, and M.T. Orchard. Joint space-frequency segmentation using balanced wavelet packet trees for least-cost image representation. *IEEE Transactions on Image Processing*, 6(9):1213–1230, 1997.

[66] G.K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):31–43, 1991.

[67] A. Leger, T. Omachi, and G.K. Wallace. JPEG still picture compression algorithm. *Optical Engineering*, 30(7):947–954, 1991.

[68] D. LeGall. MPEG: A video compression standard for multimedia applications. *Communications of the ACM*, 34(4):47–58, 1991.

[69] ITU-Telecommunication Standarization Sector. Draft recommendation H.263, 1994.

[70] A. Gersho. Asymptotically optimal block quantization. *IEEE Transactions on Information Theory*, 25(4):373–380, 1979.

[71] W.R. Bennett. Spectra of quantized signals. *Bell Syst. Tech. J.*, 27:446–472, 1948.

[72] J.J.Y. Huang and P.M. Schultheiss. Block quantization of correlated gaussian random variables. *IEEE Transactions on Communications Systems*, 11(3):289–296, 1963.

[73] B. Macq. Weighted optimum bit allocations to orthogonal transforms for picture coding. *IEEE Journal on Selected Areas in Communications*, 10(5):875–883, 1992.

[74] A.J. Ahumada and H.A. Peterson. Luminance-model-based DCT quantization for color image compression. volume 1666 of *Proceedings of the SPIE*, pages 365–374, 1992.

[75] S. Daly. Application of a noise-adaptive Contrast Sensitivity Function to image data compression. *Optical Engineering*, 29(8):977–987, 1990.

[76] A.B. Watson. DCT quantization matrices visually optimized for individual images. In B.E. Rogowitz, editor, *Human Vision, Visual Processing and Digital Display IV*, volume 1913, 1993.

[77] B. Girod. Motion compensation: Visual aspects, accuracy and fundamental limits. In M.I. Sezan and R.L. Lagendijk, editors, *Motion Analysis and Image Sequence Processing*, 1993.

[78] J.M. Foley. Human luminance pattern mechanisms: Masking experiments require a new model. *Journal of the Optical Society of America A*, 11(6):1710–1719, 1994.

[79] J.G. Daugman. Uncertainity relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America*, 2(7):1160–1169, 1985.

[80] J. Fdez-Valdivia, J.A. García, J. Mtnez-Baena, and X.R. Fdez-Vidal. The selection of natural scales for images using adaptive Gabor filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):458–469, 1998.

[81] X.R. Fdez-Vidal, J.A. García, J. Fdez-Valdivia, and R. Rdguez-Sánchez. The role of integral features for perceiving image distortion. *Pattern Recognition Letters*, 18(8):733–740, 1997.

[82] G.E Legge. A power law for contrast discrimination. *Vision Research*, 18:68–91, 1981.

[83] A.M. Pons. *Estudio de las Funciones de Respuesta al Contraste del Sistema Visual*. PhD thesis, Dpt. d'Òptica, Facultat de Física, Universitat de València, Julio 1997.

[84] A.B. Watson. Efficiency of a model human image code. *Journal of Optical Society of America A*, 4(12):2401–2417, 1987.

[85] G.E Legge and J.M. Foley. Contrast masking in human vision. *Journal of the Optical Society of America*, 70:1458–1471, 1980.

[86] A.M. Pons, J. Malo, J.M. Artigas, and P. Capilla. Image quality metric based on multidimensional contrast perception models. *Displays*, Accepted Feb. 1999.

[87] F.W. Campbell and J.G. Robson. Application of Fourier analysis to the visibility of gratings. *Journal of Physiology*, 197:551–566, 1968.

[88] D.H. Kelly. Receptive field like functions inferred from large area psychophysical measurements. *Vision Research*, 25(12):1895–1900, 1985.

[89] N.B. Nill. A visual model weighted cosine transform for image compression and quality assessment. *IEEE Transactions on Communications*, 33:551–557, 1985.

[90] L.A. Saghri, P.S. Cheatheam, and A. Habibi. Image quality measure based on a human visual system model. *Optical Engineering*, 28(7):813–819, 1989.

[91] R. Rosenholtz and A.B. Watson. Perceptual adaptive JPEG coding. *Proceedings IEEE Intl. Conf. on Image Processing*, pages 901–904, 1996.

[92] F. Dufaux and F. Moscheni. Segmentation-based motion estimation for second generation video coding techniques. In L. Torres and M. Kunt, editors, *Video Coding: A Second Generation Approach*, 1996.

[93] E. Reusens et al. Dynamic approach to visual data compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(1):197–211, 1997.

[94] ISO/IEC JTC1/SC29/WG11 N1909. Overview of the MPEG-4 version 1 standard, 1997.

[95] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

[96] S.S. Beauchemin and J.L. Barron. The computation of optical flow. *ACM Computer Surveys*, 27(3):433–467, 1995.

[97] J.K. Aggarwal and N. Nandhakumar. On the computation of motion from a sequence of images: A review. *Proceedings IEEE*, 76(8):917–935, 1988.

[98] G.M. Schuster and A.K. Katsaggelos. A video compression scheme with optimal bit allocation among segmentation, motion and residual error. *IEEE Transactions on Image Processing*, 6(11):1487–1502, 1997.

[99] A.H. Barr. Superquadrics and angle-preserving transformations. *IEEE Computer Graphics & Applications*, 1(1):11–23, 1981.

[100] K.N. Nygan, K.S. Leong, and H. Singh. Adaptive cosine transform coding of images in the perceptual domain. *IEEE Transactions on Acoustics Speech and Signal Processing*, 37(11):1743–1750, 1989.

[101] K.N. Nygan, H.C. Koh, and W.C. Wong. Hybrid image coding scheme incorporating human visual system characteristics. *Optical Engineering*, 30(7):940–946, 1991.

[102] N.B. Nill and B.R. Bouzas. Objective image quality measure derived from digital image power spectra. *Optical Engineering*, 32(4):813–825, 1992.

[103] J.O. Limb. Distortion criteria of the human viewer. *IEEE Transactions on Systems, Man & Cybernetics*, 9(12):778–793, 1979.

[104] H. Marmolin. Subjective MSE measurements. *IEEE Transactions on Systems, Man & Cybernetics*, 16(3):486–489, 1986.

[105] P.J.G. Barten. Evaluation of subjective image quality with the square root integral method. *Journal of the Optical Society of America A*, 7(10):2024–2031, 1990.

[106] P.J.G. Barten. Evaluation of the effect of noise on subjective image quality. In *Human Vision, Visual Processing and Digital Display II*, volume 1453 of *Proceedings of the SPIE*, 1991.

[107] D.J. Granrath. The role of human visual models in image processing. *Proceedings of the IEEE*, 69(5):552–561, 1981.

[108] D.R Fuhrmann, J.A. Baro, and J.R. Cox. Experimental evaluation of psychophisical distortion metrics for JPEG-encoded images. volume 1913 of *Proceedings of the SPIE*, pages 179–190, 1993.

[109] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[110] R.M. Gray. Vector quantization. *IEEE Acoustics Speech & Signal Processing Magazine*, pages 4–28, April 1984.

[111] A. Segall. Bit allocation and encoding for vector sources. *IEEE Transactions on Information Theory*, 22:162–169, 1976.

[112] D.H. Kelly. Motion and vision II: Stabilized spatiotemporal threshold surface. *Journal of the Optical Society of America*, 69(10):1340–1349, 1979.

[113] D.H. Kelly. Spatiotemporal variation of chomatic and achromatic contrast thresholds. *Journal of the Optical Society of America A*, 73(6):742–749, 1983.

[114] B.L. Beard, S. Klein, and T. Carney. Motion thresholds can be predicted from contrast discrimination. *Journal of the Optical Society of America A*, 14(9):2449–2470, 1997.

[115] E. Martinez-Uriegas. Color detection and color contrast discrimination thresholds. In *Proceedings of the OSA Annual Meeting ILS–XIII*, page 81, Los Angeles, 1997.

[116] F.A.A. Kingdom and P. Whittle. Contrast discrimination at high contrast reveals the influence of local light adaptation on contrast processing. *Vision Research*, 36(6):817–829, 1996.

[117] J. Malo. Caracterización numérica del sistema visual como filtro lineal en un dominio de Gabor. Technical report, Dpt. d'Òptica, Universitat de València, Junio 1995.

[118] D. Gabor. Theory of communication. *J. Inst. Elect. Eng.*, 93:429–457, 1946.

[119] J. Malo, A. Felipe, M.J. Luque, and J.M. Artigas. On the intrinsic two-dimensionality of the CSF and its measurement. *Journal of Optics*, 25(3):93–103, 1994.

[120] T. Ebrahimi and M. Kunt. Image compression by Gabor expansion. *Optical Engineering*, 30(7):873–880, 1991.

[121] W.H. Press, B.P. Flannery, S.A. Teulosky, and W.T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, 1992.

[122] C.K. Chui. *An Introduction to Wavelets*. Academic Press, New York, 1992.

[123] B. Macq and H.Q. Shi. Perceptually weighted vector quantization in the DCT domain. *Electronics Letters*, 29(15):1382–1384, 1993.

[124] A. Nicoulin, M. Mattavelli, W. Li, A. Basso, A.C. Popat, and M. Kunt. Image sequence coding using motion compensated subband decomposition. In M.I. Sezan and R.L. Lagendijk, editors, *Motion Analysis and Image Sequance Processing*, pages 226–256. Kluwer Academic Publishers, 1993.

[125] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Massachussets, 1996.

[126] J. Malo, F. Ferri, J. Albert, and J. Soret. Comparison of perceptually uniform quantization with average error minimization in image transform coding. *Electronics Letters*, 35(13):1067–1068, 1999.

[127] M. Bierling. Displacement estimation by hierarchical block-matching. *Proceedings of the SPIE, Conf. Visual Communications and Image Processing*, 1001:942–951, 1988.

[128] M.H. Chan, Y.B. Yu, and A.G. Constantinides. Variable size bock matching motion compensation with applications to video coding. *Procedings IEE, Vision Image and Signal Processing*, 137(4):205–212, 1990.

[129] F. Dufaux, I. Moccagatta, B. Rouchouze, T. Ebrahimi, and M. Kunt. Motion-compensated generic coding of video based on multiresolution data structure. *Optical Engineering*, 32(7):1559–1570, 1993.

[130] J. Li and X. Lin. Sequantial image coding based on multiresolution tree architecture. *Electronics Letters*, 29(17):1545–1547, 1993.

[131] V. Seferidis and M. Ghanbari. Generalised block-matching motion estimation using quad-tree structured spatial decomposition. *Proceedings IEE, Vision Image and Signal Processing*, 141(6):446–452, 1994.

[132] M.H. Lee and G. Crebbin. Image sequence coding using quadtree-based block-matching motion estimation and classified vector quantization. *Proceedings IEE, Vision Image and Signal Processing*, 141(6):453–460, 1994.

[133] R.M. Gray, P.C. Cosman, and K.L. Oehler. *Incorporating Visual Factors into Quantizers for Image Compression*. In A.B. Watson, editor, *Digital Images and Human Vision*, pages 61–138, Massachusetts, 1993. MIT Press.

[134] P.C. Cosman, R.M. Gray, and R.A. Olshen. Evaluating quality of compressed medical images: SNR, subjective rating and diagnostic accuracy. *Proceedings of the IEEE*, 82(4):919–932, 1994.

[135] A.B. Watson. *Perceptual Aspects of Image Coding*. In A.B. Watson, editor, *Digital Images and Human Vision*, pages 61–138, Massachusetts, 1993. MIT Press.

[136] H. Gish and J.N. Pierce. Asymptotically efficient quantizing. *IEEE Transactions on Information Theory*, 14:676–683, 1968.

[137] Y. Yamada, S. Tazaki, and R.M. Gray. Asymptotic performance of block quantizers with difference distortion measures. *IEEE Transactions on Information Theory*, 26(1):6–14, 1980.

[138] P.A. Chou, T. Lookabaugh, and R.M. Gray. Entropy-constrained vector quantization. *IEEE Transctions on Acoustics Speech and Signal Processing*, 37(1):31–42, 1989.

[139] M. Lightstone and S.K. Mitra. Image-adaptive vector quantization in an entropy-constrained framework. *IEEE Transactions on Image Processing*, 6(3):441–450, 1997.

[140] J. Max. Quantizing for minimum distortion. *IEEE Transactions on Information Theory*, 6(1):7–8, 1960.

[141] S.P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):127–135, 1982.

[142] D.L. McLaren and D.T. Nguyen. Removal of subjective redundancy from DCT-coded images. *Proceedings IEE-I*, 138(5):345–350, 1991.

[143] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson. Measurement of scene-dependent quality variations in digitally coded television pictures. *IEE Proceedings on Vision, Image and Signal Processing*, 142(3):149–154, 1995.

[144] D.W. Lin, M.H. Wang, and J.J. Chen. Optimal delayed coding of video sequences subject to a buffer size constraint. *Proceedings of the SPIE, Conf. Visual Communications and Image Processing*, 2094(0):223–234, 1993.

[145] D. Kersten. Predictability and redundancy of natural images. *Journal of the Optical Society of America A*, 4(12):2395–2400, 1987.

[146] D.C. Knill, D. Field, and D. Kersten. Human discrimination of fractal images. *Journal of the Optical Society of America A*, 7(6):1113–1123, 1990.

[147] T.S. Huang and A.N. Netravali. Motion and structure from feature correspondences: A review. *Proceedings IEEE*, 82(2):252–268, 1994.

[148] J.R. Jain and A.K. Jain. Displacement measurement and its applications in interframe image coding. *IEEE Transcations on Communications*, 29(12):1799–1808, 1981.

[149] R. Koch. Dynamic 3D scene analysis through synthesis feedback control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:556–568, 1993.

[150] IEEE Transactions on Circuits and Systems for Video Technology. Special issue on MPEG-4, Feb. 1997.

[151] Signal Processing: Image Communications. Special issue on MPEG-4, July 1997.

[152] H. Li, A. Lundmark, and R. Forchheimer. Image sequence coding at very low bit rates: A review. *IEEE Transactions on Image Processing*, 3(5):589–609, 1994.

[153] J.L. Barron and R. Eagleson. Recursive estimation of time-variying motion and structure parameters. *Pattern Recognition*, 29:797–818, 1996.

[154] F. Moscheni, F. Dufaux, and H. Nicolas. Entropy criterion for optimal bit allocation between motion and prediction error information. *Proceedings of the SPIE, Conf. Visual Communications and Image Processing*, 2094:235–242, 1993.

[155] L. Torres and M. Kunt. *Video Coding: A Second Generation Approach.* Kluwer Academic Publishers, Boston, 1996.

Agradecimientos (por diferentes motivos...)

Gracias a mis directores por haber soportado estoicamente a un doctorando tan cabezota como este. Cualesquiera otros hace tiempo que me habrían mandado a la mierda.

Gracias a mis amig@s por haber soportado estoicamente a un doctorando tan cabezota como este. Cualesquiera otr@s hace tiempo que me habrían mandado a la mierda.

Gracias, de verdad, a Cristina por haber soportado estoicamente a un doctorando tan cabezota como este. Seguro que cualquier otra hace tiempo que me habría mandado a la mierda.