Establecimiento de una métrica para medir la diferencia de imágenes

Daniel Hernández

Resumen—Se presenta una nueva métrica basada en [1], añadiendo la asimilación de luminancia y efectos de contraste que incluyen atributos psicofísicos y psicológicos conocidos del sistema visual del primate, con la contribución del modelo de CIWaM de Xavier Otazu et al [2], [3], que son las suposiciones relacionadas con la frecuencia espacial y el contraste del entorno. Las predicciones del modelo se compararon con experimentos psicofísicos de discriminación [4], En primer lugar, se entrenó con la secuencia en blanco y negro, con diferentes reglas que se adaptan a los datos. Después se analizó el desempeño en las secuencias de color y finalmente se probó cómo predice la imagen que cambia y la correlación con las opiniones subjetivas de observadores de la base de datos CSIQ (sólo distorsión JPEG) [7]. En general, el modelo produjo buenas predicciones en las secuencias de blanco y negro y en las secuencias de color, aunque podrían mejorarse y mostró una correlación muy alta con las opiniones subjetivas de CSIQ, por lo que podría usarse para determinar cómo de diferentes parecen dos imágenes similares para observadores humanos.

Palabras clave— Inducción de color, Inducción de luminancia, Métrica de diferencia de imagen, Sistema visual humano, Testing psicofísico, Transformada wavelet

Abstract—A new metric is presented based on a previous metric [1], we add brightness assimilation and constrast effects that include known psychophysical and physiological attributes of the primate visual system, with the contribution of CIWaM model Xavier Otazu [2], [3], they are assumptions related to the spatial frequency and the contrast surround energy. The model's predictions are compared to psychophysical discrimination experiments [4], the model is trained with black and White sequence, using some rules that improve fitting, then we analise the performance on color sequences and finally it's tested against the CSIQ database (only JPEG distortion) [7], taking into account discrimination prediction and DMOS correlation. In general, the model produces good predictions on black and White sequences and color sequences, although they could be improved and has a very high correlation to CSIQ subjective opinions, therefore it could be used to decide how different two similar images look to human observers.

Index Terms— Brightness induction, Color induction, Human Visual System, Image difference metric, Psychophysical testing, Wavelet transform

1 Introducción

LOS modelos de comparación de imágenes toman como entrada la imagen original y la distorsionada y su salida es la predicción de la calidad visual de la imagen distorsionada respecto a la original. La efectividad de los modelos, se comprueba examinando cómo se ajusta a los diferentes datos experimentales. [7]

Los primeros métodos se basaban en la energía de las distorsiones, como Mean-Squared Error (MSE) y Peak Signal to Noise Ratio (PSNR) que operan en la diferencia de los píxels. Root-mean-squared (rms) es otro ejemplo en el que la energía de las distorsiones se mide con la luminancia. Más recientemente, se desarrollaron métodos basados en las propiedades del sistema visual humano. [7]

La mayoría de métodos basados en el sistema visual humano (SVH), funcionan con descomposición perceptual que imita el análisis local de frecuencia espacial que lleva a cabo la visión. Esta descomposición es típicamente seguida por etapas de procesado que tienen en cuenta propiedades psicofísicas como la sensibilidad del contraste y visual masking.

Otra clase de métodos que se han propuesto recientemente [7] no modelan explícitamente las etapas de la visión, sino que operan en principios globales de lo que el SVH intenta conseguir cuando ve una imagen distorsionada. Estos modelos suelen incluir alguna forma de extracción estructural o de información, asumen que una imagen de alta calidad incluye buena parte de esta información (fronteras de objetos y/o regiones de alta entropía).

El objetivo de este trabajo es obtener una nueva métrica de comparación de imágenes similares, que tenga en cuenta el SVH, en concreto la inducción de luminancia y

- *E-mail de contacte: dhernandez0@gmail.com*
- *Menció realitzada: Computació*
- *Treball tutoritzat per: C. Alejandro Párraga (CVC)*
- *Curs 2013/1*

color, para ello se basa en [1], [2] y [3]. Además se intentará ajustar el modelo para un caso real, la comprensión de imágenes JPEG.

2 MODELO

2.1 Introducción

El término luminancia se refiere a la percepción no cuantitativa de luz obtenida por la luminosidad de una referencia visual [11]. Esta luminancia no solo depende de la luz que recibe la retina desde la referencia visual, también en la distribución espacial de la luz y sus alrededores [2].

La inducción de luminancia se refiere al cambio de apariencia debido a la luz de alrededor y sus efectos se clasifican de acuerdo a la dirección perceptual del cambio. Cuando el cambio en luminancia de la referencia visual se aleja de la luminancia de su entorno, se le llama contraste de luminancia [12] y cuando el cambio va hacia el entorno, se llama asimilación de luminancia [2]. El contraste de entorno es una medida de inducción de luminancia, que tiene en cuenta la relación entre el contraste de cada punto y su entorno, para determinar cuanta inducción hay.

BIWaM es un modelo de inducción de luminancia de bajo nivel [2], combina frecuencia espacial, orientación espacial y contraste del entorno para explicar la asimilación/contraste de luminancia. Esto se lleva a cabo mediante una descomposición wavelet multiresolución que separa la imagen de entrada en componentes de diferentes frecuencias espaciales y orientación.

Se suelen usar funciones Gabor para modelar las respuestas de las áreas visuales corticales [13], (ya que las neuronas reaccionan más a ciertas frecuencias espaciales que a otras [1]), las cuales son buenas descripciones de los primeros campos receptivos del sistema visual pero no es una función invertible, así que no puede recuperarse la imagen original, necesaria para CIWaM. El wavelet, no es una función Gabor estrictamente, pero es parecida [2].

La recuperación de la luminancia perceptual de la imagen se hace dando un peso a cada coeficiente de cada plano wavelet con una CSF modificada que tiene en cuenta el contraste de entorno. La distancia de observación también se tiene en cuenta en la CSF modificada. Esta función se explicará en la sección 2.4.

La extensión de este modelo al espacio de color es CIWaM [3]. Usa el espacio de color propuesto por MacLeod y Boynton [8] que refleja las excitaciones relativas de los fotoreceptores humanos. También está relacionado directamente con la fisiología de la ruta metabólica visual del primate y al córtex en términos de señales oponentes de color postreceptorales [14]. CIWaM procesa cada canal de color por separado.

Otro modelo en el que nos vamos a basar es [1], su objetivo es determinar si dos imágenes muy similares son diferentes para los observadores. Para hacerlo, compara el contraste dentro de cada canal de frecuencia espacial, diseñado para que tenga la frecuencia de células simples en el córtex visual (entre 1 y 1,5 octaves [15], [16], [17], [18]). El contraste [19] se define dentro de una banda de frecuencia F en cada punto (x, y) de la imagen (1).

$$C_F(x,y) = \frac{a_F(x,y)}{l_m(x,y)} \tag{1}$$

En la ecuación (1), el numerador es un filtro pasa banda de la imagen convolucinada con un operador simétrico circular (2).

$$A_F = e^{-\left(\frac{(f-F)^2}{2\sigma^2}\right)} \tag{2}$$

Y el denominador de (1), es una estimación de la media local de luminancia, derivado de un filtro pasa bajo de la imagen usando un operador circular simétrico (3).

$$L_F(f) = e^{-\left(\frac{f^2}{2\sigma^2}\right)} \tag{3}$$

En (2) y (3), f es la frecuencia espacial, F es la frecuencia central y σ es la desviación estándar de las curvas de frecuencia Gausiana. Esta definición de contraste se ajusta al principio de que es modulación de luminancia dividido entre la media de luminancia y captura el contraste *percibido* de imágenes complejas [19].

Después el modelo, utiliza la suma de Minkowski con dos exponentes diferentes para modelar la interacción entre los canales y entre los receptores dentro de cada canal.

Muchos investigadores usan la distancia Euclidea o la de Minkowski (1) en sus modelos [10].

$$Q = \sqrt[E]{\sum_j \left[p_{0,j} - p_{1,j} \right]^E} \tag{4}$$

En (4), p es un píxel, su índice en la imagen es j, 0 y 1 son las imágenes a comparar.

Los estudios iniciales [20], [21], sobre la relación entre estímulos de bajo contraste y los canales de frecuencia espacial han mostrado la necesidad de representar interacciones no lineales entre los diferentes canales en los modelos. Uno de ellos es el modelo de suma probabilística (*probability summation model*) [21], que aplica un gran número de canales independientes al estímulo inicial, además también se añade ruido (modelado en forma de suma probabilística Gausiana) al estímulo para representar la variabilidad de las respuestas de los sujetos. Una versión computacionalmente más simple llamada modelo de magnitud de vector (*vector magnitude model*) que combina la salida de todos los canales haciendo una suma (no euclidea), antes de sumar el ruido y tomar la decisión de detección [22].

En (4), la distancia Euclidea sería E = 2, la suma de Min-

kowski que es aproximadamente la suma probabilística, sería E = 4, y para encontrar la diferencia máxima absoluta sería E -> infinito.

El modelo también incluye la función sensibilidad de contraste para grantings sinusoidales (CSF), mediante la función dipper (se explica en la sección 2.4), para tener en cuenta que el contraste depende de la frecuencia espacial para el observador.

Nuestro modelo es una modificación de [1], sustituyendo filtros de Fourier por planos wavelet y sustituyendo el contraste por el contraste de entorno [2], [3]. Para que el modelo tenga en cuenta la inducción de luminancia y color se modifica la función CSF.

Creemos que mejorará respecto a [1], porque detectará características de inducción que el anterior no detectaba.

2.2 Descomposición en planos wavelet

El primer paso es quitar la corrección gamma en caso de que la imagen la tenga y después se pasa al espacio propuesto por MacLeod y Boynton [8].

La imagen se descompone en diferentes planos wavelet que representan diferentes frecuencias (0,5, 2, 8, 32, 128 y 512 cpd o cycles per degree), con las que se simulan la suposición 1 de BIWaM [2]. Esta suposición es que la asimilación de luminancia sólo sucede si cada punto de la imagen y su alrededor tienen una frecuencia espacial similar dentro del rango de cerca de un octave.

La suposición 2 de BIWaM [2], dice que la asimilación de luminancia es mayor cuando el centro y su alrededor tienen la misma orientación. Esta no la vamos a aplicar ya que buscamos simplicidad y tener en cuenta la orientación complica más las cosas, además las imágenes con las que vamos a probarlo no tiene mucha influencia la orientación.

De forma que la imagen se puede reconstruir con estos planos wavelet, como vemos en (5)

$$I = \sum_{s=1}^{n} w_s + c_n \tag{5}$$

En (5), I es la imagen original, w_s es un plano wavelet de la imagen, n representa el total de planos y c_n la imagen no representada por los planos wavelet.

2.3 Contraste del entorno

Una vez tenemos la imagen dividida en planos wavelet, queremos simular la suposición 3, que dice, cuando el contraste del entorno aumenta, la asimilación de luminancia también y viceversa [2].

Para implementar la inducción de luminosidad necesitamos tener en cuenta el contraste del entorno respecto al centro y usamos el mismo método que CIWaM y BIWaM [2], [3], pero sin tener en cuenta la orientación.

$$\sigma_{cen}^2 / \sigma_{sur}^2 \tag{6}$$

Definimos (6) dónde dividimos las desviaciones estándar de los coeficientes del wavelet del centro y entorno respectivamente, de manera que representan la interacción centro-entorno para cada punto (x, y).

Las regiones para calcular la desviación se modelaron con cuadrados de 5x5 para el centro (para incluir dos periodos Nyquist completos) y 13x13 para el entorno (casi tres veces más grande que la región del centro, ratio aproximado que sugieren en [23] y [24] y se mide psicofísicamente en [25]). Aunque ambas regiones se centran en el mismo punto, la del entorno no incluye los puntos de la central.

Un estudio [26] muestra que no existen linealidades en el contraste en los canales de frecuencia espacial. Esto fue modelado con una función similar a la de [27], que tuvo éxito en reproducir las respuestas de las neuronas corticales V1 [28], [29], [30]. Por lo que se introduce (7), de igual forma que se hace en BIWaM [2].

$$Z_{ctr} = \frac{r^2}{1 + r^2} \tag{7}$$

En (7), el contraste local decrece cuando el del entorno crece y viceversa, por lo tanto Zctr puede verse como una estimación no lineal del grado de luminosidad inducido por el entorno en el centro.

2.4 Extended CSF

La función CSF representa el threshold de detección para grantings dependiendo sólo de la frecuencia [6]. La modificación que hace la ECSF es incluir la contribución del contraste del entorno, de forma que cuando el contraste del entorno se incrementa, la asimilación (el contraste del centro disminuye) se incrementa también y viceversa como se puede ver en la ilustración 1.

La función CSF extendida, de CIWaM [3], tiene este nombre porque toma como valores Zctr (en el modelo final tomará la media de Zctr de las dos imágenes), frecuencia y *sthr* (frecuencia donde la CSF alcanza el pico).

En el caso de Zctr = 1 (cuando no hay desbalance entre el contraste centro-entorno) es la función CSF. Por lo tanto, la CSF se ve modificada por la influencia del contraste del entorno respecto al centro y *sthr* se usa para tener en cuenta la distancia desde la que se ve la imagen. Para los canales de color se usa una ECSF diferente (la misma para los dos canales), ya que la CSF es diferente para los canales de color, porque las características de transferencia espacial de los canales cromáticos son diferentes del canal acromático (el último es paso banda en frecuencia espa-

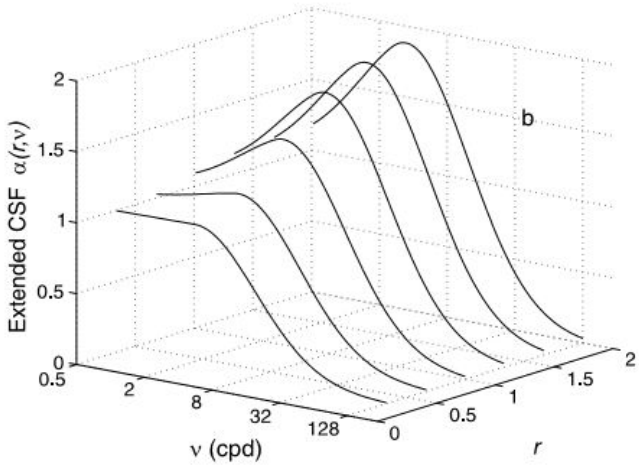


Ilustración 1. Representación gráfica de algunos valores de r de la ECSF para el canal de luminancia. La CSF corresponde al caso particular r = 1 [3].

cial y el primero es paso bajo) [6]. Entonces, la ECSF representa el threshold a partir del cual es posible detectar un cambio para cada píxel.

Hasta aquí hemos implementado el modelo de CIWaM quitándole la orientación. El thresold que nos da la ECSF es el que usaremos para ajustar la función dipper y así conectarlo con la otra parte del modelo que está basado en [1].

2.5 Dipper

Para determinar si la diferencia de contraste del entorno entre dos imágenes en una banda de frecuencia es suficientemente grande para ser detectada, necesitamos saber cómo el observador puede discriminar cambios en el contraste de Michelson de grantings sinusoidales (8).

$$C = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}} \tag{8}$$

En (8), $L_{\max}$ y $L_{\min}$ son el máximo y mínimo valor de luminancia en el granting. El numerador, representa la modulación de luminancia y el denominador es la media de luminancia de granting.

La función dipper se produjo haciendo la media de las medidas psicofísicas de 3 observadores [1]. Esta función toma el contraste de Michelson y representa la diferencia de contraste a partir de la cual el observador nota un cambio. Como se puede observar en la ilustración 2, es un template que debe ser multiplicado por el valor de la CSF de la frecuencia en la que estemos trabajando, los datos fueron normalizados.

En nuestro caso, para cada píxel y banda de frecuencia, calcularemos la media de contraste entre las dos imágenes y ese será nuestro *contraste de referencia* de la función dipper, la *diferencia de contraste* la calcularemos igual, para cada píxel y banda de frecuencia, restamos los contrastes.
La función dipper, es un template y debe multiplicarse por el valor que retorna la ECSF para representar la función dipper para la frecuencia que estemos usando y añadir la influencia de la inducción de luminancia.

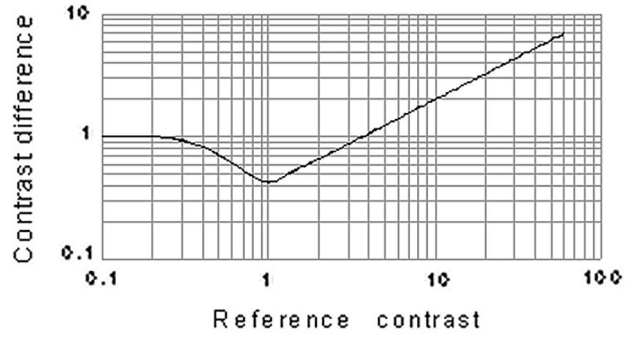


Ilustración 2. Template estándar de discriminacion de contraste "dipper". La forma se obtuvo con observadores dicriminando entre grantings con constraste de Michelson ligeramente diferentes. El gráfico representa la diferencia de contraste desde la que es discriminable la diferencia entre un par de grantings pintados contra el contraste de referencia en coordenadas log-log [1].

Si la diferencia que nos da es mayor al valor que nos da la dipper con nuestro *contraste de referencia*, el cambio en ese píxel debería ser visible, ya que la función dipper representa la diferencia de contraste a partir de la cual es posible percibir (están en el límite perceptual) un cambio para un contraste y frecuencia determinado.

2.6 Modelo para buscar el threshold

Tal como se hace en [1], calculamos el valor absoluto de la diferencia dentro del recuadro no modificado por la máscara 206 x 206 píxels. Pero en nuestro caso, en vez de restar contraste, restaremos el contraste del entorno Zctr. F es la frecuencia, i y j son las imágenes que estamos comparando.

$$\Delta Z_F(x,y) = \left[Z_{F,i}(x,y) - Z_{F,\mathrm{j}}(x,y) \right] \tag{9}$$

En la ecuación (10), calculamos la media de Zctr para cada par de píxels.

$$\overline{Z_F}(x,y) = \frac{Z_{F,i}(x,y) + Z_{F,j}(x,y)}{2} \tag{10}$$

El siguiente paso es comparar el valor de la diferencia con el correspondiente punto de la función dipper, para saber si ese punto tiene una diferencia visible entre ambas imágenes.

$$I_F(x,y) = \frac{\Delta Z_F(x,y)}{D_F(\overline{Z_F}(x,y))} \tag{11}$$

En (11), D es la función dipper, e I es el ratio entre la diferencia y la dipper. En vez de calcular la diferencia en log, se hace así por simplicidad computacional y tiene un

efecto de dar más peso que están muy por encima de la dipper.

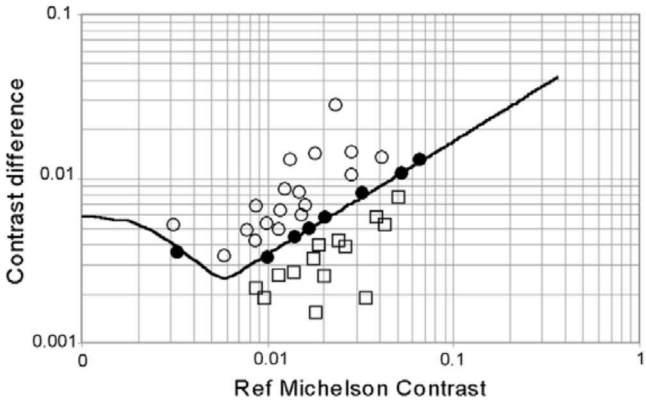


Ilustración 3. Función dipper, los puntos representan los valores de ΔZ_F pintado contra su medio valor $\overline{Z_F}$. Los puntos abiertos son valores suprathreshold (fácilmente detectable por el observador). Los puntos rellenos son valores discriminables (justo en el threshold) y los cuadrados son valores que no son discriminables [1].

En la ilustración 3, vemos un ejemplo de función dipper. Los puntos que están por debajo de la dipper no son percibibles por el observador. En cambio, los puntos sobre la dipper y por encima representan cambios que están en el límite perceptual (podrían percibirse).

Como en [1], usamos la suma de Minkowski (12) con exponente beta = 4 (aproximado a la suma probabilística) que modela la interacción entre píxeles y planos wavelet. Esto contempla la posibilidad de que todos los píxels interaccionan dando más peso a los puntos que pasan la dipper.

(12) $$Output = (\sum input^{\beta})^{\frac{1}{\beta}}$$

Aplicando la suma de Minkowski hacemos la media de los ratios de toda la imagen en (13).

(13) $$\overline{I_F} = \sqrt[4]{\frac{\sum_{x,y} I_F(x,y)^4}{numpixels}}$$

De esta forma, los puntos que están por encima de la dipper, aumentan su distancia de esta. En (13), numpixels es el número de píxels que tiene la imagen. Los puntos que sean 1 estarán sobre la dipper y los superiores a 1, por encima de ella, por lo tanto una media de 1, es un plano wavelet con cambios detectables.

Con la ecuación (14), sumamos el efecto en cada plano wavelet.

(14) $$jnd = \sqrt[4]{\sum_F \overline{I_F^4}}$$

Si dos planos wavelet dan media de detectables, es más probable que la imagen tenga cambios detectables por el observador, por eso se usa la suma de Minkowski.

Con la ecuación (15), adaptamos esto a las imágenes de color, sumando los tres canales con el mismo peso.

(15) $$jnd = \frac{jnd_1 + jnd_2 + jnd_3}{3}$$

Este número final (*just noticable difference)* tiene que ser superior a 1 para considerar que las imágenes son suficientemente diferentes en la regla más simple (regla A), se discutirá en la sección 2.8.

2.7 Experimentos de secuencias morph

Para medir y ajustar el modelo se usaron las secuencias de morph de Lovell et al [4]. Es un limón que se va convirtiendose en un pimiento rojo gradualmente, el fondo son hojas y la secuencia cuenta con 40 imágenes más la inicial en el experimento con tres observadores. Se estimó el porcentaje del morph en el que cada observador era capaz de diferenciar la imagen inicial de la secuencia.

Además, el morph está modificado con varios filtros, en total hay 49 secuencias morph. Las imágenes se transformaron al espacio de frecuencia con Fourier y entonces se aplicaron los filtros parecidos a blur o sharpen tanto en luminancia como en color, como vemos en (16).

(16) $$weight(f) = f^{-\alpha}$$

En (16), f es la frecuencia espacial y alpha es el parámetro slope.

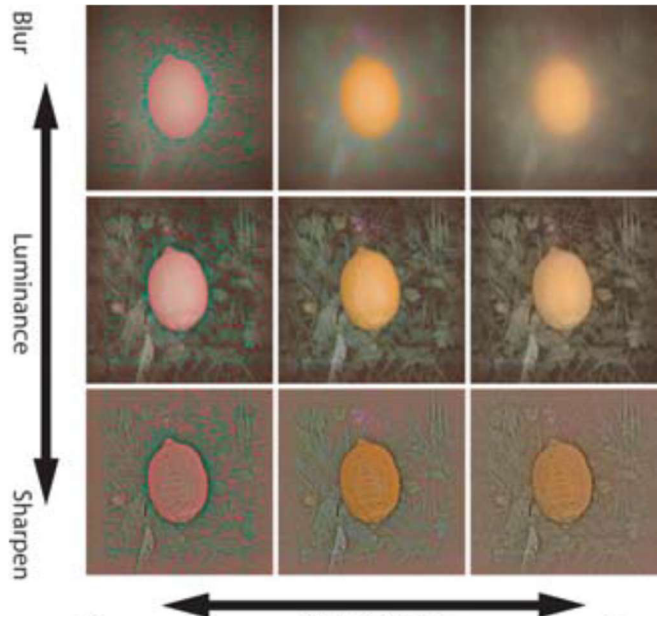


Ilustración 4. Ejemplos de imágenes manipuladas. Al limón (imagen central) se le aplica efecto blur o sharpen separadamente en los canales de color o luminancia. A la imagen superior izquierda se le ha aplicado el efecto blur en el canal de luminancia y en los dos de color, a la inferior izquierda se le aplicó el efecto sharpen en los tres canales.

El experimento se concentra en establecer los thresholds de discriminación, donde las diferencias entre las imágenes son relativamente sutiles, para poder optimizar el modelo con estos datos después. Como se puede apreciar en la ilustración 4, se aplicaron los valores alpha (+0,4, +0,8 y +1,2) que hacen efecto blur y una reducción de las frecuencias altas y (-0,4, -0,8 y -1,2) que hacen efecto sharpening e incremento de frecuencias altas. El valor 0 deja las imágenes como las originales, de esta forma, entrenamos el modelo para imágenes en las que diferentes bandas de frecuencia tienen importancia.

Todos estos efectos, se aplican tanto en los canales de color (en ambos igual) como en el de luminancia. También hay otra secuencia con las imágenes en blanco y negro, por lo tanto sólo fue modificado el canal de luminancia.

2.8 Reglas de modelado

Para probar y ajustar el modelo, cargamos cada una de las secuencias y medimos cual es la imagen en la que el modelo determina que es percibible un cambio para compararlo con los datos psicofísicos. Estas reglas están basadas en las del modelo [1].

Regla A:
Usamos beta con valor 4, por lo que estamos aplicando la suma probabilística. Si jnd es superior o igual a 1, la imagen tiene cambios detectables por el observador. No tiene parámetros libres. El porcentaje se obtiene haciendo interpolación lineal con el par j y j+1 (j+1 es la imagen en la que se detecta que es suficientemente grande el cambio).

Regla B:
Usamos beta con valor 4 también y sólo detecta que hay suficientes cambios cuando el jnd es superior o igual a 1. Tiene un parámetro libre, es una constante que se aplica a los jnd de cada plano wavelet para que se ajuste a los datos experimentales (sólo imágenes en blanco y negro). Usamos la función fminbnd de MATLAB y buscamos entre -0,5 y 0,5.

Regla C:
Usamos beta con valor 4, por lo que estamos aplicando la suma probabilística. Se coge el jnd de la imagen en la que se comienzan a detectar cambios según los experimentos para alpha = 0 (blanco y negro) y se utiliza este valor como "magic number" para determinar si una imagen tiene cambios visibles o no en las otras secuencias.

Regla D:
Usamos beta con valor 4, por lo que estamos aplicando la suma probabilística. Primero buscamos el jnd óptimo para la secuencia de imágenes en blanco y negro, usando la función fminbnd de MATLAB buscando el valor entre 1 y 3, una vez tenemos el "magic number", buscamos la constante para ajustar los jnds de todos los planos wavelets tal como hicimos en la regla B.

En las reglas que necesitan iterar para conseguir el mejor valor para algún parámetro libre ajustándose a los datos experimentales, hemos usado la función fminbnd y (17) para obtener la distancia de los datos reales. La ecuación (17) es como Mean Square Error pero no hay la necesidad de dividirlo entre el número de elementos.

(17)
$$SSE_R = \sum_{i=1}^{n} (\overline{Y}_i - Y_i)^2$$

En (17), Y es el dato experimental y media de Y es el dato que predice el modelo. Sólo se tiene en cuenta la secuencia en blanco y negro para ajustar el modelo por complejidad computacional. El modelo ideal daría SSE = 0, ya que cada punto de la curva pasaría exactamente por los datos experimentales. En el apéndice 1, tenemos un resumen del modelo.

3 APLICACIÓN CON DISTORSIÓN JPEG

El funcionamiento de nuestra nueva métrica fue comprobado en una base de datos comúnmente usada con este tipo de modelos (CSIQ [7]), consiste en 30 imágenes con 6 tipos de distorsión y de 4 a 5 niveles de distorsión, nosotros usaremos sólo la distorsión JPEG que incluye 5 niveles de distorsión (comprimido con más pérdida).

Ilustración 5. Por orden de izquierda a derecha, imagen 1600 original, tercera distorsión JPEG, quinta distorsión JPEG

En la ilustración 5, se puede ver la original, la tercera distorsión y la quinta de la imagen "1600".

Las imágenes son evaluadas subjetivamente basado en desplazamiento lineal de cuatro monitores 1920 x 1200 LCD calibrados donde se visualizaban las imágenes, uno al lado del otro y a la misma distancia del observador, 70 cm.

Todas las versiones de distorsión se mostraron a la vez ordenadas en orden de distorsión. La base de datos se basa en 5.000 evaluaciones de 25 observadores en forma de DMOS. Los observadores tenían una edad entre 21 y 35 años y una visión normal o correcta.

Para evaluar cómo se comporta el modelo con respecto a las medidas DMOS, primero se le aplica una transformación logística a las predicciones de la métrica, que ponen los valores predichos en la misma escala que los de DMOS. Y para conseguir una relación lineal entre las predicciones y las opiniones DMOS, usamos la transformación logística (18) recomendada por Video Quality Experts Group [9].

(18)
$$f(x) = \frac{t_1 - t_2}{1 + e^{\frac{x - t_3}{t_4}}} + t_2$$

En (18), los parámetros t se eligen para minimizar el MSE entre los valores predichos y DMOS. Estamos usando la misma transformación que en [7] para poder comparar con las distintas métricas.

Para medir el desempeño de la métrica, usaremos la medida (19) Spearman Rank-order Correlation Coefficient (SROCC) para medir cómo de bien se correlaciona la predicción con el DMOS y cómo predice el orden relativo de las imágenes distorsionadas. Usaremos esta medida para comparar el modelo con otros en [7].

(19)
$$p = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

En (19), d es la distancia entre la predicción y DMOS y n es el número de elementos.

(20)
$$R_{out} = \frac{N_{false}}{N_{total}}$$

También usaremos el outlier ratio (20), donde Nfalse es el número predicciones fuera de dos desviaciones estándar y Ntotal el total de elementos. Se usaron dos desviaciones estándar porque tienen el 95% de probabilidad de contener el punto dentro. Esta métrica es útil porque tiene en cuenta que cuando la desviación es grande, acercarse a la media no es importante ya que los observadores no coinciden mucho y también el caso contrario.

Además de analizar la correlación de las predicciones con DMOS, analizaremos que DMOS tienen las imágenes que el modelo detecta como cambiadas, para ver si estas imágenes tienen un DMOS parecido. Si esto es así, querría decir que el modelo detecta cambios coincidiendo con la opinión de los observadores.

4 METODOLOGÍA

El código está basado en la versión MATLAB de CIWaM [3], modificado usando la función dipper [1] e implementando todo lo demás necesario para programar el modelo que hemos definido anteriormente.

Con todo implementado, fue un proceso iterativo, probando todas las reglas y creando reglas nuevas hasta encontrar las que mejor siguieran el patrón U experimental y tuviesen un error bajo para la secuencia de blanco y negro.

Con todas las reglas, sacamos el mejor modelo para la secuencia de blanco y negro y lo aplicamos a las secuencias de color para ver cómo se comportaba.

Finalmente usando los datos psicofísicos JPEG de CSIQ [7], para analizar el desempeño de esta métrica en una aplicación real, por brevedad sólo explicamos los resultados de la regla más óptima. Este último paso es importante porque era importante ver cómo funcionaba con unas imágenes reales y un problema real cómo el presentado.

5 RESULTADOS

5.1 Resultados de secuencias de blanco y negro

Los resultados los analizaremos primero respecto a la secuencia en blanco y negro y después pasaremos a analizar el comportamiento del modelo respecto a las secuencias de color.

En la ilustración 6, se muestran los resultados de la regla A (azul), y los resultados experimentales de los observadores KB (rojo), HS (verde) y CAP (lila).

Como podemos apreciar, el modelo tiene la misma forma en U que los datos experimentales, está más cerca del observador HS en general, con KB también tiene un buen comportamiento excepto en +1,2 que para nuestro modelo es bastante más sencillo encontrar la diferencia tanto para HS como para KB. El modelo tiene un error MSE de 389.

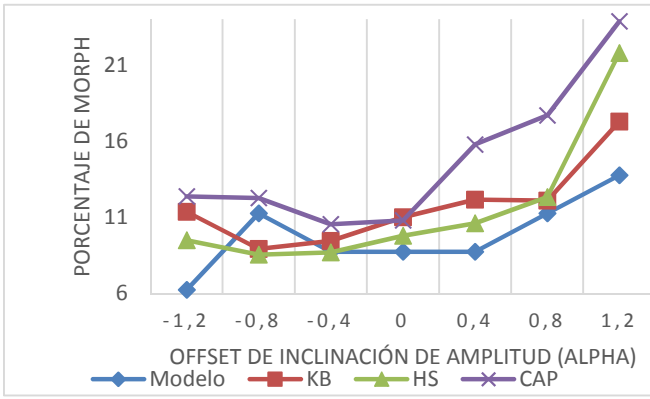


Ilustración 6. Resultados de la regla A

Para CAP, en el punto del centro, -0,4 y -0,8, se ajusta bien el modelo, pero en la parte positiva vemos un incremento de la dificultad más alto que para el modelo, aun así, en general el modelo encaja en la forma de U de las otras funciones.

Es llamativo que la secuencia con shapening más pronunciado (-1,2) sea la más fácil de discriminar según el modelo, pues es la única que no sigue la forma de la función en U, debería ser más difícil o igual de discriminar que (-0,8 y -0,4).

En la imagen de la secuencia -1,2, que se detecta como cambiada respecto a la original, el plano wavelet que más influencia tiene para ser detectada es el de frecuencia más baja, teniendo en cuenta que el sharpening es incrementar las frecuencias altas. Esto lleva a pensar que quizás se le está dando demasiada importancia a las frecuencias bajas.

Una de las mejoras posibles sería conseguir que este punto hiciera el comportamiento normal respecto a los demás y formara una U, dando menos importancia a las frecuencias bajas. Aún así, el modelo se ajusta bastante bien a los datos.

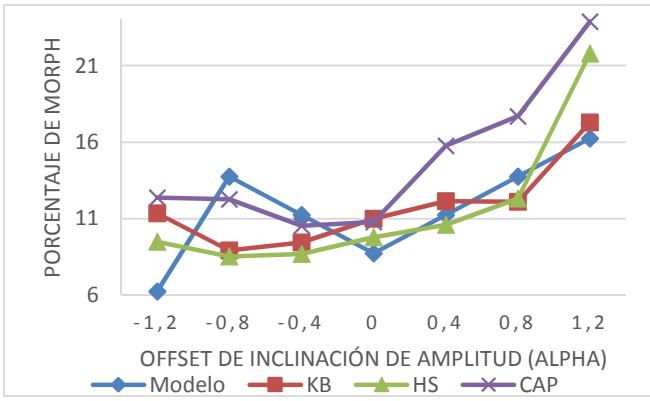


Ilustración 7. Resultados de la regla B

En la ilustración 7, se muestran los resultados de la regla B (azul), y los resultados experimentales de los observadores KB (rojo), HS (verde) y CAP (lila).

Este modelo cambia respecto al anterior en tener un parámetro libre para ajustarse a los resultados experimentales, es una constante que se suma a los jnds de cada planos wavelet. El efecto respecto a la regla A es un aumento en todos los puntos menos -1,2 (máximo sharpening) y 0 (sin modificar).

La constante que minimiza el error es -0,354 y tiene un error MSE de 279, es el 71,72 % de la regla anterior. La parte positiva del offset es la que más beneficiada se ve por el cambio, ajustándose mejor especialmente a HS. Por otro lado en la parte negativa al subir se ajusta bien a CAP pero se aleja de HS y KB, el efecto es que el error respecto a CAP baja y el de HS y KB sube y el error total baja. El punto -1,2 sigue dando los mismos problemas que en la regla A.

Por lo que en resumen, tenemos una modificación que mejora el error total gracias a una constante que hace que el cambio se detecte más tarde en casi todas las secuencias pero no consigue mejorar el caso más problemático, el sharpening de -1,2.

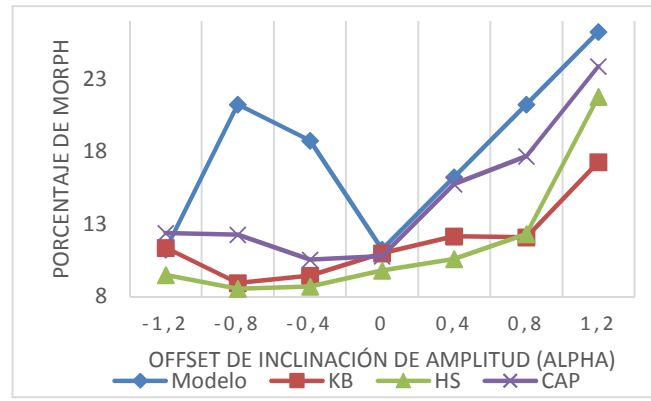


Ilustración 8. Resultados regla C

En la ilustración 8, se muestran los resultados de la regla C (azul), y los resultados experimentales de los observadores KB (rojo), HS (verde) y CAP (lila).

Este modelo se basa en coger el jnd de alpha = 0 y usarlo como "magic number" para determinar cuándo las otras secuencias tienen cambios que se pueden percibir. El error MSE de esta versión alcanza 986, notablemente mayor que las otras dos reglas, en concreto es 3,54 veces el error de la regla B, hasta ahora la mejor.

Respecto a la regla B, vemos un aumento en todos los puntos, en la parte positiva de la gráfica (blur) vemos que están incluso por encima de CAP, con lo cual el error respecto a los otros observadores aumentará, aun así, está dentro de lo aceptable. En cambio, en la parte negativa de la gráfica (shapening) podemos apreciar que los resultados del modelo están muy distantes de los experimentales, excepto el -1,2 que nos ha dado el mejor resultado de todas las reglas hasta ahora. Por lo tanto, el peor modelo hasta ahora es este especialmente por la parte negativa de la gráfica.

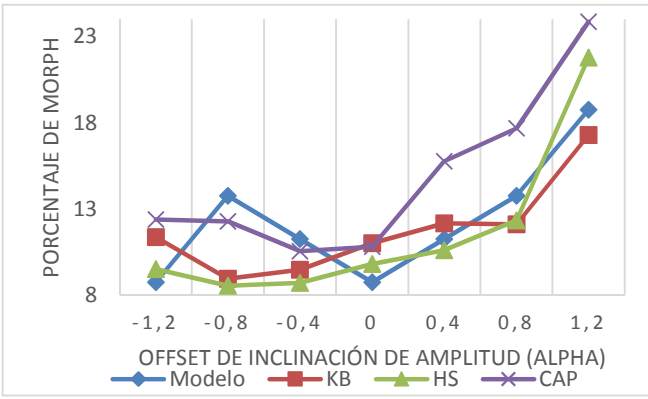


Ilustración 9. Resultados regla D

En la ilustración 9, se muestran los resultados de la regla D (azul), y los resultados experimentales de los observadores KB (rojo), HS (verde) y CAP (lila).

Esta regla busca un "magic number" para determinar si la imagen ha cambiado suficiente, de forma que se ajuste a los datos experimentales lo mejor posible y después busca la constante a sumar al jnd de cada plano wavelet que también se ajusta mejor a los datos.

La constante que minimiza el error es 0,0072 y el magic number es 1,4722 esta regla tiene un error MSE de 173, un 62% de la regla B, por lo tanto la regla con error más bajo.

La clave de esta regla para tener un error tan bajo es que siempre está cerca de alguno de los observadores, en la parte positiva de la gráfica, está cerca de HS y KB, entre medio de los dos y en la negativa cerca de CAP menos en el punto -1,2 que tiene un comportamiento extraño como en las otras reglas y coincide con HS.

Respecto a la regla B, vemos que en las modificaciones más extremas de las imágenes (1,2 y -1,2) se aumentan la dificultad y para las demás secuencias tenemos un resultado similar al de B.

La métrica se ajustó bastante bien a las secuencias de blanco y negro, excepto la regla C, las otras tres se ajustan bien a las formas de las curvas de los tres observadores.

5.2 Resultados de secuencias de color

Una vez hemos analizado las secuencias de blanco y negro, vamos a pasar a ver cómo se comportan estas reglas del modelo con las secuencias de color.

TABLA 1
ERROR DE LAS SECUENCIAS DE COLOR

	CAP	HS	KB	Total
Regla A	732,53	184,82	257,11	1.174,46
Regla B	368,09	578,30	525,58	1.471,96
Regla C	2.761,57	5.379,66	4.966,13	13.107,36
Regla D	298,57	773,10	690,09	1.761,76

Como podemos apreciar en la tabla 1, la regla C da unos resultados mucho peores que el resto de reglas, por lo tanto la podemos descartar para que no dificulte la interpretación de la gráfica.

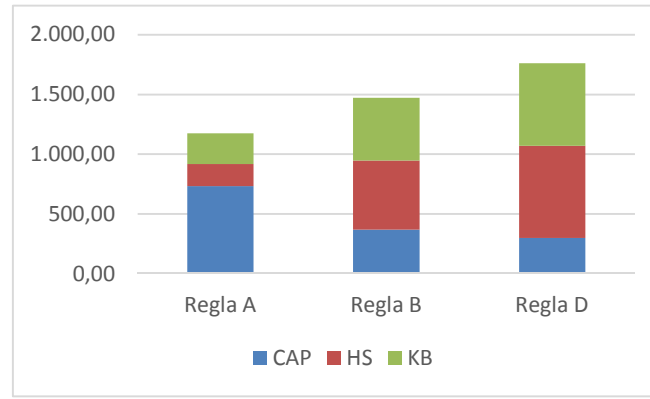


Ilustración 10. Resultados de las reglas A, B y D para las secuencias de color

Como podemos ver en la ilustración 10, vemos que la regla A es la más óptima y la regla D la menos óptima. Es llamativo que los resultados con las secuencias de color son justo al revés que las secuencias en blanco y negro, lo cual podría deberse a que al ajustar el modelo a los resultados estamos haciendo que dependan mucho de estos y no generalicen bien.

Otra cosa a tener en cuenta, es que la influencia de los tres canales no tiene por qué ser equivalente y es posible que buscar los pesos óptimos para cada canal hubiera mejorado el desempeño, porque no sabemos qué canal es el que hace discriminar cada imagen (en el modelo se usa 0,33 para cada canal).

Otro aspecto a destacar es que el error de A se debe sobre todo al error respecto al observador CAP. Más de la mitad del error se debe a la distancia a CAP, mientras que parece ajustarse muy bien a los otros dos observadores. De igual forma, la regla D, se adapta mucho mejor a CAP que a HS y KB. Quizás una combinación de las dos reglas o usar una curva por observador y luego ponderarlas hubiera mejorado los resultados.

5.3 Resultados calidad JPEG

Como explicamos en la sección 3, primero vamos a comparar el SROCC que obtenemos con las imágenes con distorsión JPEG de la base de datos CSIQ [7]. Por brevedad, sólo vamos a aplicar la regla D, que es la que mejor resultados nos ha dado.

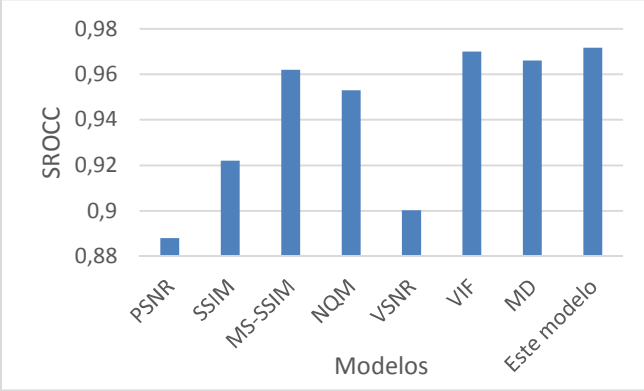


Ilustración 11. Resultados SROCC de diferentes métricas, el mejor es el nuestro con 0,9717

Como vemos en la ilustración 11, nuestro modelo es el que tiene la correlación más alta de los comparados con SROCC, concretamente 0,9717.

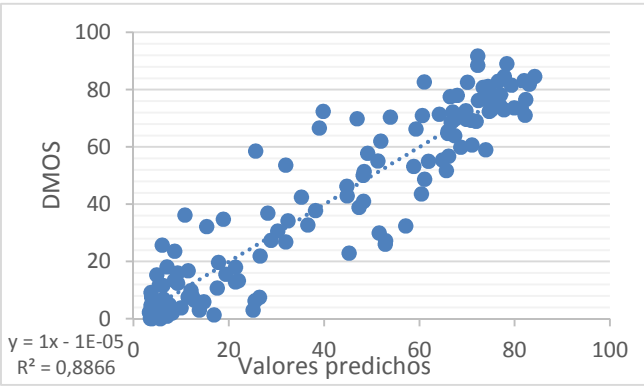


Ilustración 12. Gráfico de correlación de valores predichos y DMOS, como vemos hay una correlación fuerte

En la ilustración 12, apreciamos que hay una correlación fuerte entre DMOS y los valores predichos por la métrica. Además tenemos un Rout de 31,33%, por lo que el 68,67% de los puntos entran dentro de sigma por 2, el área con probabilidad de 95% en la distribución normal. Por lo tanto, parece según los datos, que los valores que nos devuelve la métrica (jnd de cada imagen) están correlacionados con los valores DMOS experimentales.

Pero nuestro modelo dice en qué imagen hay un cambio suficientemente grande para que el observador lo note. Para comprobar esto, vamos a analizar si los valores DMOS donde se detecta suficiente cambio están juntos y son valores pequeños (mucha calidad respecto al original). Si esto se cumple, el modelo estaría prediciendo bien.

En la ilustración 13, tenemos la campana de Gauss de los valores DMOS correspondientes a las imágenes que se han detectado como cambiadas.

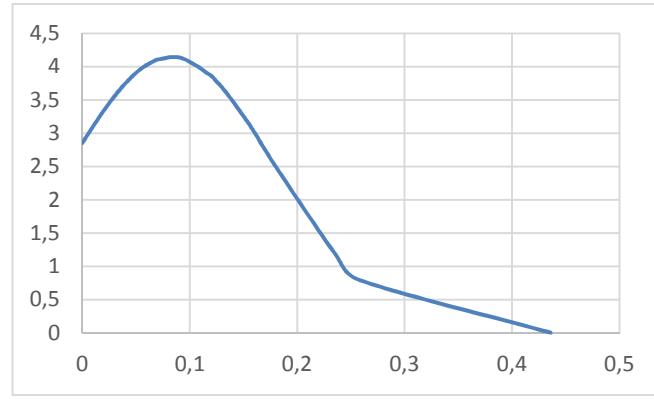


Ilustración 11. Distribución normal de DMOS detectados por la regla D

En general, alrededor de 0,1 suele detectar la imagen como cambiada. En este sentido es bastante intuitivo, pues un DMOS bajo significa que los observadores vieron pocos cambios y todas las imágenes tienen un DMOS parecido, cosa que hace pensar que la detección está funcionando bien, detecta la imagen cambiada con el mismo valor DMOS.

TABLA 2
VALORES DMOS DONDE SE DETECTA LA IMAGEN CAMBIADA

	DMOS anterior	DMOS detectado
Child swim.	0,030	0,236
Couple	0,034	0,258
Sunset color	0,031	0,436

En la tabla 2, comprobamos que los valores más altos de DMOS tienen el problema de dar un salto muy grande entre la anterior y en la que es detectada. Los tres tienen menos de 0,035 en la anterior a la detectada y en la siguiente imagen pasa a ser un valor alto, por lo tanto fácilmente detectable, lo que demuestra que estos valores son casos especiales por la naturaleza de la imagen, que hace que al subir la compresión se degrade muy rápidamente la calidad y no da lugar a una imagen de calidad intermedia.

6 CONCLUSIÓN

A pesar del problema con la imagen de sharpening más pronunciado (-1,2), las reglas logran ajustarse bien a los datos experimentales de las secuencias en blanco y negro, con las secuencias de color también se logra un ajuste bueno.

Los resultados con la base de datos CSIQ son muy positivos, más teniendo en cuenta que el modelo no se consiguió ajustar al 100% en las pruebas de las secuencias de blanco y negro y color. Quizás este problema no existe con las pruebas JPEG, porque las imágenes de la secuencia -1,2 no son imágenes naturales, por lo que al probar el modelo en una base de datos de degradación JPEG no interfiere, al ser todas las imágenes de la base de datos naturales.

Las posibles mejores serían, sobre todo, intentar ajustar el modelo a la secuencia -1,2, probando con menos importancia para las frecuencias bajas. Otra posible mejora sería buscar constantes para cada uno de los tres canales que ajusten mejor el modelo a las secuencias de color. Y finalmente, se podría probar de hacer una regla diferente para cada observador y luego intentar ponderarlas u obtener resultados por separados, ya que las que se ajustan a CAP no pueden ajustarse del todo bien a HS y KB y viceversa.

AGRADECIMIENTOS

Me gustaría agradecer a mi tutor de prácticas por toda la ayuda prestada y tiempo dedicado en ayudarme a entender y desarrollar el proyecto.

También quiero agradecer a Marina por su paciencia y apoyo.

BIBLIOGRAFÍA

[1] C. Alejandro Párraga, "Is the human visual system optimised for encoding the statistical information of natural scenes?", PhD disertation, Department of Experimental Psychology, Faculty of Science, 2003.

[2] Xavier Otazu, Maria Vanrell and C. Alejandro Párraga, "Multiresolution wavelet framework models brightness induction effects", Vision Research. 48, 733-751, 2008.

[3] Xavier Otazu, Maria Vanrell y C. Alejandro Párraga, "Toward a unified chromatic induction model", Journal of Vision, 10(12):5, 1–24, 2010.

[4] P. George Lovell, C. Alejandro Párraga, Tom Troscianko, Caterina Ripamonti y David J. Tolhust, "Evaluation of a Multiscale Color Model for Visual Difference Prediction", ACM Transactions on Applied Perception, 3, pp. 155-178, 2006.

[5] A. J. Ahumada, Jr., "Computational Image Quality Metrics: A Review", SID Digest. 305-8, 1993.

[6] Mullen K.T, "The contrast sensitivity of human color vision to red-green and blue-yellow chromatic gratings", Journal of Physiology-London. 359, 381-400, 1985.

[7] E. C. Larson y D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," Journal of Electronic Imaging, 19 (1), 2010.

[8] Donal I. A. MacLeod y Robert M. Boynton, "Chromaticity diagram showing cone excitation by stimuli of equal luminance", Department of Psychology, University of California, 1978.

[9] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii," 2003.

[10] Huib de Ridder, "Minkowski-metrics as a combination rule for digital-image-coding impairments", Proc. SPIE 1666, Human Vision, Visual Processing, and Digital Display III, 16, 1992.

[11] Gilchrist, A., "Seeing in Black and White", Scientific American Mind, 42-49, 2006.

[12] Heinemann, E. G., "simultaneous brightness induction as a function of inducing and test-field luminances", Journal of Experimental Psychology, 50(2), 89-96, 1955.

[13] Daugmann, J. G., "Two-dimensional spectral analysis of cortical receptive field profile", Vision Research, 20, 847-856, 1980.

[14] Derrington, A. M., Krauskpf, J., & Lennie, P., "Chromatic mechanisms in lateral geniculate nucleus of macaque", The Journal of Physiology, 357, 241-265, 1984.

[15] De Valois R.L., Albrecht D.G. y Thorell L.G, "Spatial-frequency selectivity of cells in macaque visual cortex", Vision Research. 22, 545-559, 1982.

[16] De Valois R.L., Yund E.W. y Hepler N., "The orientation and direction selectivity of cells in macaque visual cortex", Vision Research. 22, 531-544, 1982.

[17] Movshon J.A., Thompson I.D. y Tolhurst D.J., "Spatial and temporal contrast sensitivity of neurons in areas 17 and 18 of the cat's visual cortex", Journal of Physiology. 283, 101-120, 1978.

[18] Tolhurst D.J. y Thompson I.D., "On the variety of spatial-frequency selectivities shown by neurons in area-17 of the cat", Proceedings of the Royal Society of London Series B-Biological Sciences. 213, 183-199, 1981.

[19] Peli E., "Contrast in Complex Images", Journal of the Optical Society of America A. 7, 2032-2040, 1990.

[20] Robson J.G. y Graham N, "Probability summation and regional vairation in contrast sensitivity across the visual field", Vision Research. 21, 409-418, 1981.

[21] Sachs M.B., Nachmias J. y Robson J.G, "Spatial-frequency channels in human vision", Journal of the Optical Society of America A. 61, 1176-1186, 1971.

[22] Quick R.F, "A vector magnitude model of contrast detection", Kybernetik. 16, 65-67, 1974.

[23] Spitzer, H., & Semo, S., "Color constancy: A biological model and its application for still and video images", Pattern Recognition, 35(8), 1645–1659, 2002.

[24] Shapley, R., & Enroth-Cugell, C., "Visual adaptation and retinal gain controls", Progress in retinal research (Vol. 3). Oxford: Pergamon Press, 1984.

[25] Yu, C., Klein, S. A., & Levi, D. M., "Surround modulation of perceived contrast and the role of brightness induction", Journal of Vision, 1(1), 18–31, 2001.

[26] Nachmias, J., & Sansbury, R. V., "Grating contrast discrimination may be better than detection", Vision Research, 14, 1039–1042, 1974.

[27] Naka, K. I., & Rushton, W. A., "S-potentials from luminosity units in the retina of fish (cyprinidae)", Journal of Physiology, 185(3), 587–599, 1966.

[28] Albrecht, D. G., & Hamilton, D. B., "Striate cortex of monkey and cat: Contrast response function", Journal of Neurophysiology, 48(1), 217–237, 1982.

[29] Sclar, G., Maunsell, J. H., & Lennie, P., "Coding of image contrast in central visual pathways of the macaque monkey", Vision Research, 30(1), 1–10, 1990.

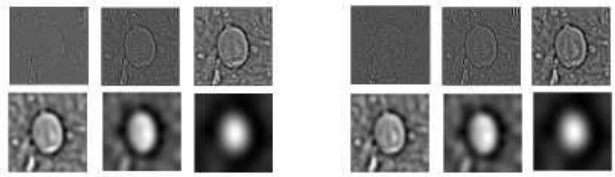
[30] Tolhurst, D. J., & Heeger, D. J., "Contrast normalization and a linear model for the directional selectivity of simple cells in cat striate cortex", Visual Neuroscience, 14(1), 19–25, 1997.

APÉNDICE

A1. RESUMEN DE LA MÉTRICA

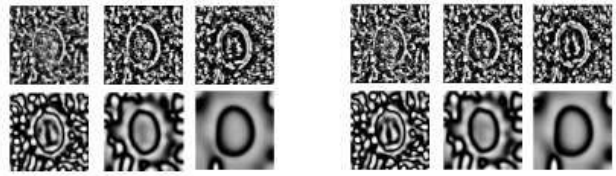
Planos wavelet

Se calculan los planos wavelet de 2, 4, 8, 16, 32 y 64 ciclos/ picture



Contraste entorno

Se calcula la desviación estándard del centro y del entorno de cada píxel y se dividen, así tenemos una medida de la influencia del contraste del centro respecto a su entorno



Diferencia de constraste de entorno

Diferencia entre contrastes de entorno para cada plano wavelet de las dos imágenes

Dipper

Se calcula la media de cada wavelet de las dos imágenes y usa la ECSF para cada píxel de esta media. Luego se pasa a la dipper y se divide la diferencia entre el valor ECSF para saber qué cambios son visibles

Decisión

Hay cuatro reglas para decidir qué imagen de la secuencia se detecta como diferente, según el número de puntos que cruzan la dipper

Regla A

Interacción entre receptores y canales, decisión basada en suma de Minkowsky.

No hay parámetros libres, cuando jnd supera 1 se entiende que la imagen es suficientemente diferente

Regla B

Interacción entre receptores y canales, decisión basada en suma de Minkowsky.

Un parámetros libre, una constante que se aplica a todos los jnd de cada wavelet para que se ajuste a los datos experimentales.

Cuando jnd supera 1 se entiende que la imagen es suficientemente diferente

Regla C

Interacción entre receptores y canales, decisión basada en suma de Minkowsky.

Cogemos el jnd de la imagen en la que se comienza a detectar cambios en la secuencia sin modificar alpha=0 y se toma como referencia para saber cuando son las imágenes detectables

Regla D

Se calcula la desviación estándard del centro y del entorno de cada píxel y se dividen, así tenemos una medida de la influencia del contraste del centro respecto a su entorno