Características principales del HVS (Human Visual System) usadas en Codificación Perceptual de Imagen y Vídeo

[latexpage]

> *Todo aquello que es perceptualmente redundante puede ser eliminado.*

Estas son las características fundamentales del sistema visual humano (HVS) que se han incorporado o tenido en cuenta en la codificación perceptual de imagen y video

En esta entrada vamos a revisar brevemente cada uno de estas características.

Color Space

La teoría tricromática del color sugiere que cualquier color puede ser descompuesto en tres canales de color. De entre los distintos espacios de color que utilizan esta teoría el RGB (Read Green Blue) es el más extendido ya que se usa directamente en en televisión y monitores, aunque ultimamente también se utilizan otros.

Con RGB cada imagen se representa en tres planos de color, donde cada uno almacena la información del color correspondiente. Uno de los inconvenientes de usar este espacio de color es que la cantidad de energía de los tres canales de color es aproximadamente la misma. Por lo que cada canal tiene aproximadamente el mismo peso y no se aprovecha

la correlación entre los tres canales.

Una alternativa es usar también tres canales, pero donde el primero lleva información de luminancia y los otros dos una combinación de dos canales de crominancia. De entre éstos, dos modelos son los más utilizados, YCbCr y YUV. Fueron diseñados para aprovechar la caracteristica del HVS que le hace más sensible a los cambios en luminancia que a los cambios de color y también con el objetivo de mantener la compatibilidad con televisiones en blanco y negro, puesto que el canal de luminancia es el único que éstos utilizaban.

Estos modelos por tanto decorrelacionan la información RGB de forma que los canales de crominancia tienen mucha menos energía, y por tanto ancho de banda (aprox. un 10%), que el de luminancia. Esta propiedad del HVS, permite además submuestrear los canales de crominancia y aplicar mayor cuantización a los mismos sin perder la información importante (luminancia) para el HVS.

Tanto el modelo YCbCr como el YUV utilizan unas matrices de conversión (Figura 1) para pasar de RGB a YCbCr o YUV y viceversa. YUV es muy usado en vídeo análogico tradicional y el YCbCr se utiliza en video digital. El canal Y es una suma pondereada del RGB y tanto UV como CbCr son diferencias ponderadas de RB con Y. Los pesos están diseñados para que UV y CbCr tiendan a cero o a un valor constante para imágenes planas monocromáticas neutras, es decir, cuando R=G=B. Por tanto la mayoría de la información se ubica en el canal Y.

$$\begin{bmatrix} Y \\ Cr \\ Cb \end{bmatrix} = \begin{bmatrix} 0.2989 & 0.5866 & 0.1145 \\ 0.5000 & -0.4183 & -0.0816 \\ -0.1687 & -0.3312 & 0.5000 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Figura 1.
Matriz de conversión de RBG
a YCbCr

La mayoría de la investigación de codificación de imagen y video se ha centrado tradicionalmente en el canal de luminancia para la proposicion de nuevos modelos pero la extensión a color es necesaria.

Las propiedades de filtrado y enmascaramiento que veremos a continuación actuan sobre la información de lumninancia del HVS.

Propiedades HVS: Filtrado/Enmascaramiento y Locales/Globales

De las tres propiedades del HVS que vamos a revisar, la sensibilidad frecuencial al contraste (frequency sensitivity) es una propiedad de filtrado, mientras que las otras dos son propiedades de enmascaramiento (masking).

El umbral de sensibilidad (*visibility threshold* T) es una medida definida como la magnitud de un estimulo en una imagen a partir de la cual el estímulo se hace visible o invisible. El estímulo puede ser cualquier señal, por ejemplo una señal sinusoidal, ruido aditivo o distorsion. En teoría, cualquier estímulo que esté por debajo del umbral de visibilidad puede ser eliminado (o si es una distorsión, tolerada o pasar desapercibida) lo que elimina redundancia perceptual.

En algunas circunstancias en vez del *visibility threshold* es necesario tener una medidad de la sensibilidad del ojo a cierto estímulo. La sensibilidad es la inversa del *visibility threshold,* es decir cuanto más alto sea el umbral menor sensibilidad tenemos y a la inversa, cuanto menor sea el umbral mayor sensibilidad tenemos.

También podemos clasificar las propiedades el HVS como Locales o Globlaes, es decir en propiedades del HVS que dependen de

características locales de la imagen y las que no.

La sensibilidad al contraste (CSF) del HVS, que se basa como veremos en la MFT del HVS puede ser considerada una propiedad global.

El Luminance Masking y el Contrast Masking se clasifican como locales puesto que dependen de la actividad (cantidad de energia) de una zona (bloque) de la imagen o de la cantidad media de luminancia de una zona (bloque) de la imagen.

CSF (Contrast Sensitivity Function)

Estudios psico-visuales han demostrado que la percepción de una distorsión depende de la respuesta en frecuencia del HVS. Se ha demostrado que el HVS actua como un filtro paso-banda con una respuesta máxima (máxima sensibilidad) en el rango de 8 cpd (cycles per degree) disminuyendo mucho para frecuencias menores y para muy altas.

La respuesta en frecuencia de un sistema viene definida por la MFT del mismo (Modulation Transfer Function). El HVS tiene una MFT como la de la Figura 2:

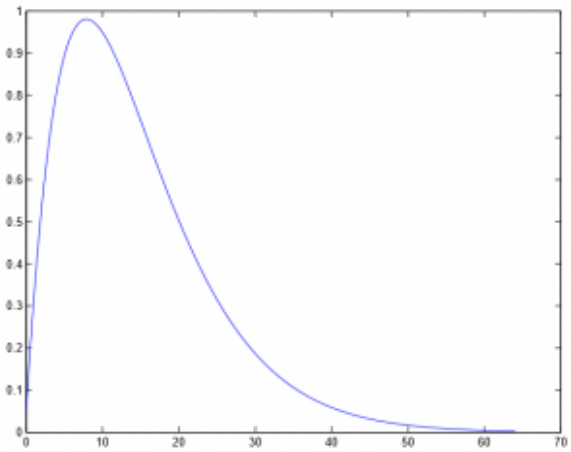

Figura 2.
Modulation Transfer Function del HVS. CSF – Contrast Sensitivity Function

En el eje Y tenemos la Sensibilidad Frecuencial al Contraste.
En el eje X tenemos la frecuencia espacial en cycles per degree (cpd)

La CSF proporciona una manera natural de incorporar criterios perceptuales en técnicas de codificación de imagen y video basadas en transformadas en frecuencia. La imagen es transformada al dominio de la frecuencia, donde cada coeficiente de un bloque (DCT – Discrete Cosine Transform) o subbanda frecuencial (DWT – Discrete Wavelet Transform) corresponde a un rango de frecuencias, para los cuales, según la curva CSF (la MFT del HVS) proporciona una sensibilidad y por tanto un umbral de detección.

Por tanto, la respuesta del HVS puede ser utilizada para modular la importancia relativa de los coeficientes transformados (aplicando pesos a los mismos) . Cuanto mayor sea el peso que le damos a un coeficiente, es decir a un rango de frecuencias, más importante lo hacemos. Por el contrario, cuanto menor peso le demos a un coeficiente menor importancia tendrá. Aquellos coeficientes ubicados en franjas frecuenciales para las cuales el HVS no tiene sensibilidad suficiente podrían ser descartados o cuantizados en mayor medida. Son aquellos coeficientes para los que necesitaríamos una cantidad de estímulo (energía) muy álta para poder percibir su presencia.

Luminance Masking

Normalmente los terminos *luminance* y *brightness* (luminancia y brillo) se usan indistintamente, aunque la luminancia es una medida física y el brillo es un descriptor subjetivo que no puede ser medido. Habitualmente el termino *grayscale* (escala de grises o nivel de gris) se refiere a la componente de

luminancia de una imagen digital. Un nivel de gris de 0 (cero) en una escala de grises de 8-bits es negro y blanco sería 255.

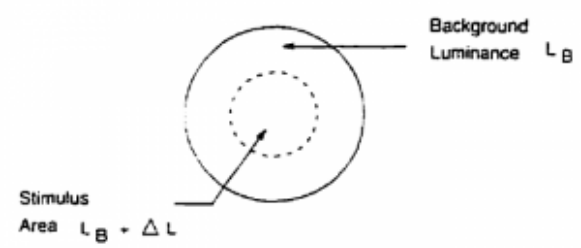

Figura 3
Icremento de luminancia en relación a la del fondo

El concepto de *luminance masking* (enmascaramiento por luminancia) se puede explicar con la Figura 3. El área contínua, que se supone el fondo (*background*) de la escena tiene una luminancia media de L_B. En el área discontínua se sitúa un estímulo basado en un incremento de la luminancia. Este estímulo es visible (se percibe el incremento de luminancia) para un valor ΔL. Experimentos demuestran que el umbral ΔL a partir del cual se detecta el incremento de luminancia es función (depende de) la luminancia del fondo, de L_B, y que se incrementa casi linealmente conforme aumenta L_B. Esto se conoce como The Weber's Law (la ley Weber).

$$
\text{The Weber's Law} \Rightarrow \frac{\Delta L}{L_B}=constant
$$

Lo que implica es que el ojo humano es menos sensible a los errores que ocurren en áreas de la escena con alta luminancia, porque ΔL es relativamente alto en estas zonas. La ley se cumple bien en zonas desde luminancia media-baja hasta alta luminancia, pero se ha reportado en otros estudios que el valor del cociente tiende a aumentar para valores muy bajos de luminancia. Es decir, para el ojo humano la sensibilidad a las distorsiones también cae en areas muy oscuras en la imagen.

Esto se puede ver en la Figura 4 que muestra el ratio de Weber.

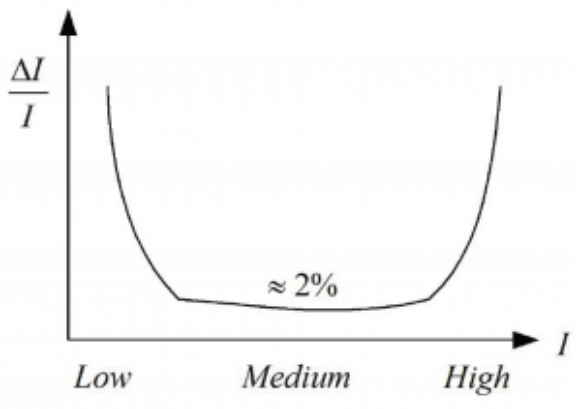

Figura 4
Contrast ratio: Weber fraction

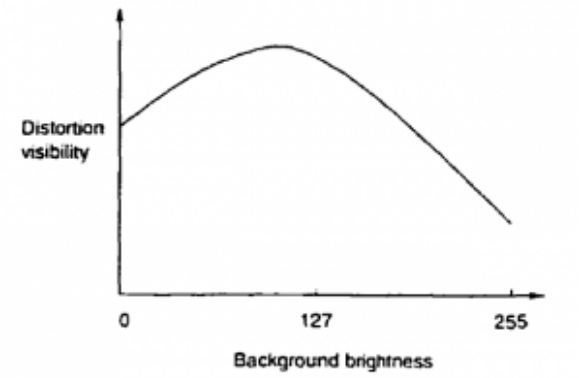

Figura 5
Sensibilidad a las distorsiones por luminancia.

La Figura 5 muestra la curva de sensibilidad a las distorsiones en relación a la luminancia. Como se puede ver la sensibilidad disminuye para valores muy oscuros de luminancia y para valores muy altos, sin embargo las distorsiones serán detectadas más fácilmente en regiones de luminancia media, puesto que ahí la sensibilidad es mayor y por tanto el umbral de detección es menor.

Vemos pues que hay zonas con menor sensiblidad a las distorsiones por luminancia, es decir con mayor capacidad de enmascarar por luminancia (muy oscuras o muy claras), de éstas decimos que tienen mayor capacidad de *luminance masking* o simplemente mayor *luminance masking*.

Esto es muy útil para la codificación de imagen y vídeo puesto que en zonas de alto luminance masking hay más redundancia perceptual y por tanto podremos introducir mayor cantidad de distorsión local.

Texture Masking

El enmascaramiento por textura o *texture masking* también se llama *spatial masking*.

En este caso la visibilidad de la distorsión disminuye cuando hay cambios grandes de la luminancia del fondo (background luminance). Es decir se cambia rápidamente (en el espacio) la luminancia (valores de gris) en un cuadro. Cuanto más cambios hay espacialmente en la luminancia un bloque o zona de la imagen podemos decir que tiene más textura. Subjetivamente parece claro que la textura enmascara u oculta ciertas distorsiones. Este efecto podemos verlo en la Figura 6, donde el mismo ruido blanco se ha añadido en la zona del cielo que en la zona de las rocas. Como vemos se percibe más el ruido en la zona del cielo, puesto que ahí hay menos textura.

Figura 6
The background image is acting as masker of a noise pattern. The original image is on the left. In the right image the noise pattern is applied to the top and bottom of the image. The texture in water and rocks makes detecting the noise pattern difficult.

En la Figura 7 se explica un ejemplo de los efectos del enmascaramiento por textura. La imagen muestra dos gráficas que representan la variación de luminancia a lo largo de una determinada orientación en la imagen. Si uno se situa al inicio del escalón, vemos cómo desde una luminancia inicial se eleva la misma en una distancia en pixels determinada (eje x). Es decir en esa orientación que representa la gráfica, la luminancia crece hasta un determinado nivel en un tenerminado número de pixels.

La gráfica superior indica una variación rápida en los niveles de luminancia en esa orientación. La gráfica inferior indica una variación lenta en los niveles de luminancia. Las líneas discontínuas representan la variación del umbral de

visibilidad ΔL que la ley de Weber indica.

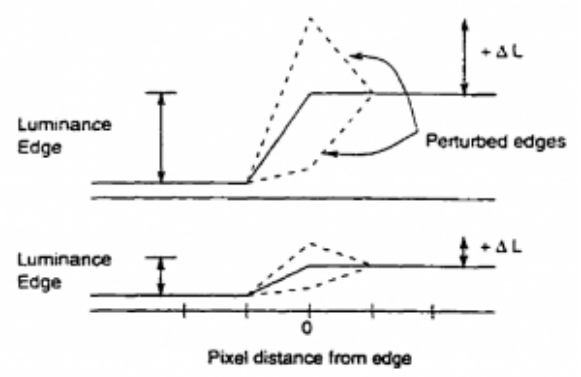

Figura 7
Efectos de Texture (spatial) Masking

Por tanto, como vemos en la Figura 7, cuando hay un estímulo presente (en este caso ΔL) cuando hay una variación de luminancia grande en las cercanías de un punto en la imagen, la variación de luminancia en esa zona, por ejemplo debido a una determinada de textura, hace que el umbral de detección suba y por tanto la sensibilidad para ver la distorsión disminuye. Así en zonas con alta textura (variación de luminancia en diversas orientaciones conjuntamente) tenemos menos posibilidad de ver las distorsiones presentes, porque la textura (variación de luminancia rápida en diferentes orientaciones), enmascara la distorsión. Cuando hay muy poca variación de luminancia, poca textura, el umbral de detección baja, con lo que la sensibilidad a detectar sube, es decir detectaremos mejor la distorsión puesto que no hay textura que enmascare.

En la codificación de imagen, el *texture masking* se ha utilizado mucho en codificaciones en el dominio del espacio como la DPCM (*Differential Pulse Code Modulation*) en la cual los bordes, donde hay altos cambios de luminancia, son detectados muy bien y por tanto en zonas con mucha textura se puede aplicar mayor cuantización. También en la transformación y codificación por bloques (DCT) se puede aplicar el *texture*

masking. Hay muchas formas de determinar la textura de un bloque. Un bloque con mucha textura implica muchos cambios de luminancia en el dominio del espacio, generándose para cada orientación de la textura una redundancia perceptual que puede ser eliminada.

Temporal Masking

El enmascaramiento temporal se basa en la propiedad de refresco del HVS. Durante la reproducción de un vídeo, cuando hay un cambio en la escena o movimiento en la misma, la visibilidad de la distorsión en la nuevas zonas afectadas por el movimiento o cambio en la escena, la visibilidad de la distorsión es baja durante una latencia breve produciendose una redundancia perceptual que puede ser explotada. Normalmente, como la componente temporal es necesaria, este enmascaramiento sólo se utiliza en vídeo.

Surgen dos tipos de enmascaramiento temporal relacionados con el tiempo, *backwards masking* y *forward masking*. Y dos tipos de enmascaramiento temporal relacionados con la posición que ocupan el estímulo y el masker, *temporal metacontrast masking* y *temporal pattern masking* respectivamente.

Temporal metacontrast masking se da cuando el target (o estímulo) y el masker no se muestran en posiciones espaciales superpuestas simultaneamente, es decir son complementarias (en la posición del estímulo no hay masker cuando el estímulo desaparece). Por otro lado el *temporal pattern masking* se da cuando tanto el másker como el target aparecen en la misma posición (los dos se muestran simultaneamente en la misma posición).

En el *fordward masking* el estímulo se mantiene un tiempo tras el cual aparece el masker. Hasta que aparece el masker hay un período de tiempo en el que el sujeto no percive el estímulo, produciendose una redundancia perceptual. El *backwards masking* ocurre cuando en un cambio brusco en la escena, la nueva

escena enmascara cierta cantidad de frames de la escena previa. Una determinada área de la escena es eliminada (preiva o posterior al frame en curso) en los niveles cognitivos (higher processing levels) del HVS. El sujeto no es consciente de perdecir correctamente estas áreas. Todavía no hay una explicación muy clara de este fenómeno, pero la más prometedora tiene que ver con la variación en la latencia de las señales neuronales del HVS en función de su intensidad.

Se han realizado estudios en compresores de vídeo donde se demuestra que la reducción transitoria de la sensibilidad del HVS es significativa en los primeros 160 ms para el *forward masking* y hasta los 200 ms para el *backward masking*. También donde se demuestra que el *backwards masking* es más significativo que el *fordward masking*.