Estudio de la
Medida de la Calidad Perceptual de Video

Ing. José Joskowicz
josej@fing.edu.uy

Índice

1. Introducción

En el presente trabajo se realiza un estudio del "estado del arte" en el ámbito de la medida objetiva de la calidad perceptual de video.

El video digital, distribuido a través de redes de comunicaciones, sufre varios tipos de distorsión durante el proceso de adquisición, compresión, procesamiento, transmisión y reproducción. Por ejemplo, las técnicas utilizadas habitualmente en la codificación digital de video introducen pérdida de información, para reducir el ancho de banda necesario para su transmisión, lo que genera distorsiones. Por otro lado, las redes de paquetes sobre las que se transporta el video (por ejemplo Internet, aunque también redes de área local LAN, las redes extendidas WAN y las redes inalámbricas), pueden introducir distorsiones adicionales, debido a las demoras, los errores y las pérdidas de paquetes, entre otros factores.

La medida objetiva de la calidad perceptual de video es necesaria para diversos tipos de aplicaciones, algunas de las cuales se describen a continuación [1]:

Monitorización
Los prestadores de servicio están interesados en conocer la calidad de las aplicaciones utilizadas por sus usuarios. Específicamente en aplicaciones de audio y video, es deseable poder medir la calidad percibida por los usuarios, a los efectos de optimizar la red. En estos casos, pueden realizarse medidas controladas, utilizando señales de prueba, y contrastando la señal recibida contra la original. Si bien esto pude realizarse con pruebas subjetivas, disponer de modelos objetivos puede sistematizar las medidas, logrando monitorizar el estado de la red en forma periódica, de manera sencilla y controlada. En este caso, los modelos utilizados pueden basarse en señales de prueba conocidas (modelos "Full Reference", como se verá más adelante)

Control de calidad y administración
Una medida objetiva y "en línea" de la calidad perceptual puede ser utilizada a los efectos de control y administración. Esta información puede ser realimentada, desde el receptor al emisor, de manera que este último tome acciones inmediatas para mantener la calidad perceptual constante (por ejemplo, aumentar o disminuir en línea el ancho de banda del video enviado). Estas técnicas deberían estimar la calidad perceptual inspeccionando únicamente la señal recibida, o eventualmente, los parámetros de la red, ya que no se dispone en este caso de la señal de referencia.

Control de Admisión y administración de recursos
Cuando los usuarios pagan por aplicaciones con cierta calidad (por ejemplo, en video a demanda en IPTV), es necesario medir la calidad percibida por todos los usuarios, y estimar como ésta se verá afectada al incluir nuevos usuarios en el sistema. Es posible que se disponga de políticas de admisión de nuevos usuarios en función del nivel de calidad percibido por otros usuarios, y/o requerido por el nuevo usuario.

Precios en base a la calidad
Algunos servicios pueden tener precios diferenciales según la calidad del mismo. En este caso, es necesario medir la calidad percibida, a los efectos de controlar que se está entregando la calidad por la que el usuario está pagando.

Nuevos Desarrollos
Las medidas objetivas de calidad percibida son una herramienta necesaria para la evaluación y desarrollo de nuevos sistemas o algoritmos. Disponer de una medida que se pueda realizar en forma automática de calidad percibida evita tener que realizar largas y costosas pruebas subjetivas, en el diseño de nuevos codificadores y decodificadores (codecs), algoritmos de realce del video, etc.

Todos estos factores hacen necesario disponer de herramientas que permitan estimar y cuantificar la calidad percibida por los usuarios en el video, de la manera más confiable posible. Un sistema ideal de medida de calidad perceptual debería dar como resultado una calificación idéntica a la que se obtendría en pruebas subjetivas promediando los resultados de un gran número de individuos

Este trabajo se organiza de la siguiente manera. En el capítulo 2 se presenta un resumen de las aplicaciones de video digital, destacando las características particulares de cada una. En el capítulo 3 se realiza un breve resumen de los sistemas de codificación digital de video, y de su forma de transmisión sobre redes IP, lo que es necesario para comprender algunos aspectos de la problemática. Los capítulos 4, 5 y 6 contienen la parte principal de este trabajo. En el capitulo 4 se introducen los conceptos de calidad perceptual, y los tipos de degradaciones que típicamente se observan en video digital. En el capítulo 5 son presentados los estudios y trabajos realizados en las técnicas de medida objetiva de la calidad perceptual de video. El capítulo 6 detalla cómo se ve afectada la calidad perceptual de video cuando éste es transmitido sobre redes IP. En base al estudio realizado, en el capítulo 7 se presentan posibles líneas de investigación en el tema. Finalmente, el capítulo 8 presenta las conclusiones.

2. Aplicaciones de video

El video es utilizado en diversos tipos de aplicaciones, las que a su vez, tienen diversos requerimientos. La TV es, quizás, la aplicación de video más conocida. Sin embargo, existen en forma cada vez más difundida un nuevo conjunto de aplicaciones de video, entre las que se encuentran la video telefonía, los servicios de video conferencia, la distribución de video a demanda a través de Internet y la IP-TV, por mencionar los más relevantes. Cada una de estas aplicaciones tiene sus características propias en lo que respecta a requerimientos de calidad, velocidades, etc.

La video telefonía es una aplicación típicamente punto a punto, con imágenes del tipo “cabeza y hombros”, y generalmente poco movimiento. Sin embargo, es una aplicación altamente interactiva, dónde los retardos punta a punta juegan un rol fundamental en la calidad conversacional percibida.

Las aplicaciones de video conferencias son típicamente punto a multi-punto. Al igual que la video telefonía, generalmente tienen poco movimiento. Además de la difusión del audio y el video es deseable en estas aplicaciones poder compartir imágenes y documentos. La interactividad también es típicamente un requisito, aunque podrían admitirse retardos punta a punta un poco mayores que en la video telefonía, ya que los participantes generalmente están dispuestos a “solicitar la palabra” en este tipo de comunicaciones.

La distribución de televisión digital, y en particular la IP-TV generan otro tipo de requerimientos. En estas aplicaciones se debe soportar todo tipo de imágenes (desde las estáticas hasta las de mayor movimiento), y la calidad percibida de la imagen juega un rol fundamental. Los usuarios de estos servicios esperan recibir la calidad por lo cual están pagando. Además de esto, otros efectos, como las demoras entre el cambio de canales (“zapping”) juega un papel importante en la experiencia del usuario. En la TV analógica, los usuarios están acostumbrados a que estos cambios de canal son prácticamente instantáneos (lo que es difícil de lograr cuando se debe, por ejemplo, tener un jittter-buffer de algunos segundos).

Finalmente es de hacer notar que en casi todas las aplicaciones de video, el audio también está presente, y juega un papel muy importante en la calidad perceptual general. La percepción del usuario respecto al video no sólo se ve condicionada por la calidad del audio, sino también por la sincronización existente entre el audio y el video. Pequeños tiempos de defasaje entre ambas señales son muy notorios (por ejemplo, al ver una persona hablando), lo que produce sensaciones

molestas, y afecta notoriamente a la calidad percibida, aún cuando la calidad de las señales de audio y de video que se estén recibiendo sean excelentes.

3. Codificación de video

Los estudios acerca de la codificación de imágenes y video comenzaron en la década de 1950. En 1984 fue introducida la estrategia de codificación utilizando la transformada discreta de coseno (DCT) [2], técnica ampliamente utilizada en los sistemas actuales de codificación. Las técnicas de compensación de movimiento aparecieron también en la década de 1980, dando origen a las tecnologías híbridas MC/DCT (Motion Compensation/Discrete Cosine Transform), utilizadas en los actuales algoritmos MPEG.

Por otra parte, las transformadas discretas de Wavelets (DWT) comenzaron también a ser utilizadas en codificación de imágenes en la década de 1980, y fueron adoptadas más recientemente dentro de las tecnologías MPEG-4 y JPEG 2000, para la codificación de imágenes fijas.

La complejidad de codificadores y decodificadores ha ido aumentando, logrando un muy alto nivel de compresión, a expensas de requerir decodificadores y, sobre todo, codificadores muy complejos, y que requieren gran capacidad de procesamiento [3]. Es de esperar que en el futuro próximo se requiera aún mayor capacidad de procesamiento, reduciendo los requerimientos de ancho de banda y mejorando la calidad percibida.

A continuación se presentan, en forma resumida, las características más destacables de las tecnologías actuales en codificación de imágenes y video, y la manera de codificar video para su transmisión sobre redes IP. No es el objetivo principal de este trabajo presentar un detalle pormenorizado de estas tecnologías, por lo que sólo se describirán brevemente sus características más relevantes.

3.1. JPEG

JPEG (Joint Photographic Experts Group) [4] es un estándar diseñado para comprimir imágenes fijas, tanto en color como en blanco y negro. El objetivo principal de este estándar fue el de lograr compresiones adecuadas, optimizando el tamaño final de los archivos comprimidos, admitiendo pérdida de calidad en la imagen. El algoritmo utilizado divide a la imagen en bloques de 8 x 8 píxeles, los que son procesados en forma independiente. Dentro de cada uno de estos bloques, se aplica la transformada discreta de coseno (DCT) bidimensional, generando para cada bloque, una matriz de 8 x 8 coeficientes. La gran ventaja de estos coeficientes, es que decrecen rápidamente en valor absoluto, lo que permite despreciar gran parte de ellos (ya que representan información de alta frecuencia espacial).
Conceptualmente, puede considerarse que cada bloque de 8 x 8 está compuesto por una suma ponderada de 64 tipos de bloques base, como se muestran en la Figura 3.1. En esta figura, cada bloque corresponde con un patrón determinado. El primer bloque (arriba a la izquierda) no tiene textura. El coeficiente asociado a este bloque se corresponde con la componente de luminancia promedio del bloque. Es conocido también como componente de DC, haciendo analogía con la "componente de continua" de una señal eléctrica. El resto de los bloques presentan patrones bien definidos, con frecuencias espaciales crecientes hacia las parte inferior-derecha de la figura.

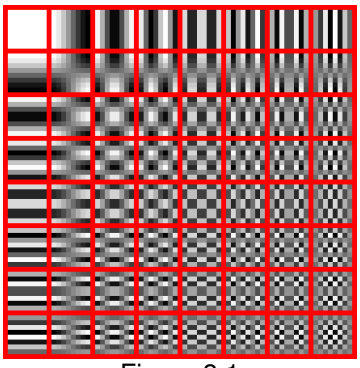


Figura 3.1

El estándar JPEG 2000 [5] está también basado en la idea de utilizar para la codificación los coeficientes de una transformación, pero en este caso se utilizan transformadas discretas de Wavelets (DWT). Esta transformada permite comprimir aún más las imágenes que la DCT. Una de las principales diferencias entre JPEG y JPEG2000 es que en esta última no es necesario dividir la imagen original en bloques. La transformada DWT se aplica a toda la imagen, lo que elimina el conocido "efecto de bloques", que se explicará en más detalle en los capítulos posteriores.

3.2. MPEG-x

MPEG-1 [6] fue originalmente diseñado por el "Moving Picture Experts Group" (MPEG) de la ISO (International Standards Organization) para el almacenamiento y reproducción digital de aplicaciones multimedia desde dispositivos CD-ROM, hasta velocidades de 1.5 Mbps. MPEG-2 [7] fue el sucesor de MPEG-1, pensado para proveer calidad de video desde la obtenida con NTSC/PAL y hasta HDTV, con velocidades de hasta 19 Mbps.

La codificación en MPEG-1 está basada en la transformada DCT para explotar las redundancias espaciales dentro de cada cuadro, y en técnicas de estimación y compensación de movimiento para explotar las redundancias temporales (entre cuadros). Las secuencias de video son primeramente divididas en "grupos de figuras" (GOP – Group of Pictures). Cada GOP puede incluir tres grupos diferentes de cuadros: I ("Intra"), P ("Predictivos) y B ("predictivos Bidireccionales"). Los cuadros del tipo I son codificados únicamente con técnicas de compresión espacial (transformada DCT dentro del propio cuadro, por ejemplo). Son utilizados como cuadros de referencia para las predicciones (hacia adelante o hacia atrás) de cuadros P o B. Los cuadros del tipo P son codificados utilizando información previa de cuadros I u otros cuadros P, en base a estimaciones y compensaciones de movimiento. Los cuadros B se predicen en base a información de cuadros anteriores (pasados) y también posteriores (futuros). El tamaño de un GOP está dado por la cantidad de cuadros existentes entre dos cuadros I. Típicamente se utilizan de 12 a 15 cuadros para un GOP, y hasta 3 cuadros entre un I y un P o entre dos P consecutivos (típicamente una señal PAL se codifica con un GOP de tamaño 12 y una NTSC con 15, amabas con no más de 2 cuadros B consecutivos). Un ejemplo tomado de [8] se muestra en la Figura 3.2 (IBBPBBPBBI), donde las flechas indican los cuadros utilizados para las predicciones. Cuando más grande el GOP, mayor compresión se puede obtener, pero a su vez existe menor inmunidad a la propagación de errores.

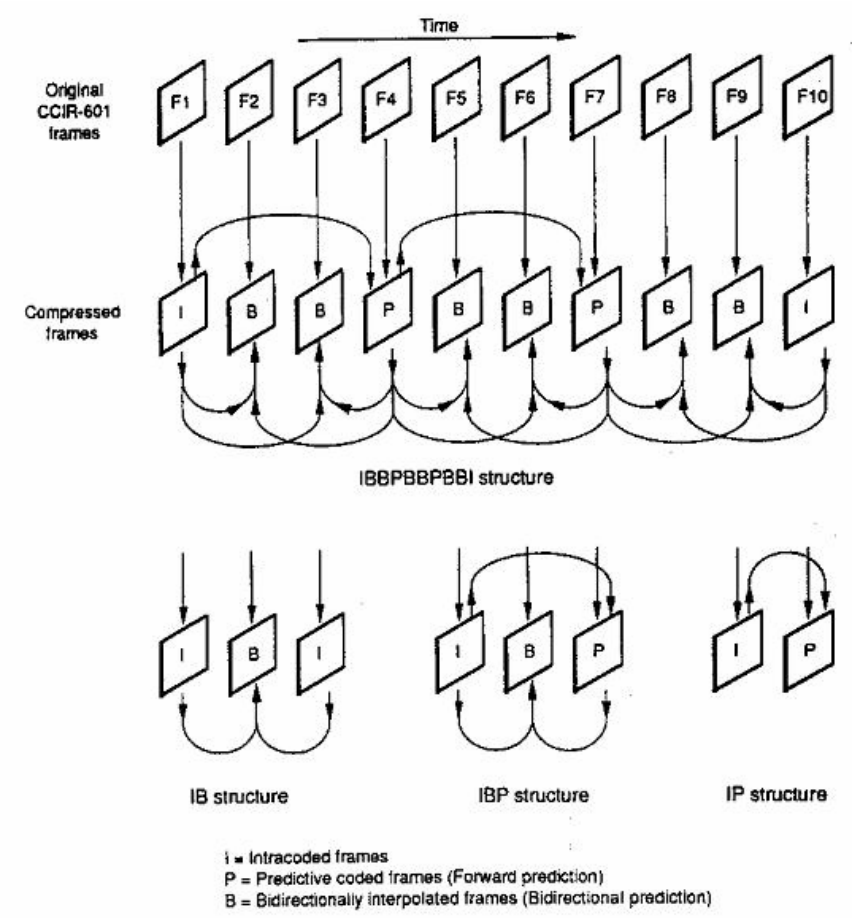


Figura 3.2

Al igual que en JPEG, en MPEG-1 se divide la imagen de cada cuadro en bloques de 8 x 8 píxeles, los que son procesados en forma independiente. Dentro de cada uno de estos bloques, se aplica la transformada discreta de coseno (DCT) bidimensional, generando para cada bloque, una matriz de 8 x 8 coeficientes. A su vez, cuatro bloques se agrupan en un "macro-bloque" de 16 x 16 píxeles, el que es utilizado como base para la estimación del movimiento. La estimación de movimiento de un macro-bloque se realiza en el codificador, comparando el macro-bloque de una imagen con todos las posibles secciones de tamaño igual al macro-bloque (dentro de un rango espacial de 512 píxeles en cada dirección) de la(s) imagen(es) siguiente(s). La comparación se realiza generalmente buscando la mínima diferencia (el mínimo valor de MSE - ver sección 5) entre el macro-bloque y la sección evaluada. Este procedimiento se basa en la hipótesis que todos los píxeles del macro-bloque tendrán por lo general un mismo desplazamiento, y por lo tanto, será más eficiente codificar un "vector de movimiento" del macro-bloque y las diferencias del macro-bloque predicho respecto del macro-bloque original. Las diferencias entre el macro-bloque predicho y el real también son transformadas mediante la DCT para su codificación.

Un flujo de video de MPEG-2 se forma de la manera descrita a continuación. Se utiliza como unidad básica un macro-bloque, compuesto típicamente por 4 bloques de luminancia y 2 de crominancia (ya que la crominancia es sub-muestreada). Los coeficientes DCT de cada uno de estos bloques son serializados, y precedidos por un cabezal de macro-bloque. Varios macrobloques contiguos (en la misma fila, y de izquierda a derecha) son agrupados formando un "slice", el que a su vez es precedido de un cabezal de "slice", el que contiene la ubicación del "slice" en la imagen y el factor de cuantización usado. Típicamente puede haber un "slice" por cada fila de macro-bloques, pero puede también haber slices con parte de una fila. Un grupo de "slices"

forma un cuadro, el que es precedido por un cabezal de cuadro, conteniendo información del mismo, como por ejemplo el tipo de cuadro (I,P,B), y las matrices de cuantización utilizadas. Varios cuadros se juntan, formando el GOP, también precedido de un cabezal de GOP. Finalmente, varios GOPs pueden serializarse en una secuencia (Elementary Stream), con su correspondiente cabezal, el que contiene información general, como el tamaño de los cuadros, y la frecuencia de cuadros. En la Figura 3.3 (tomada de [8]) se muestra un esquema del sistema de capas descrito.

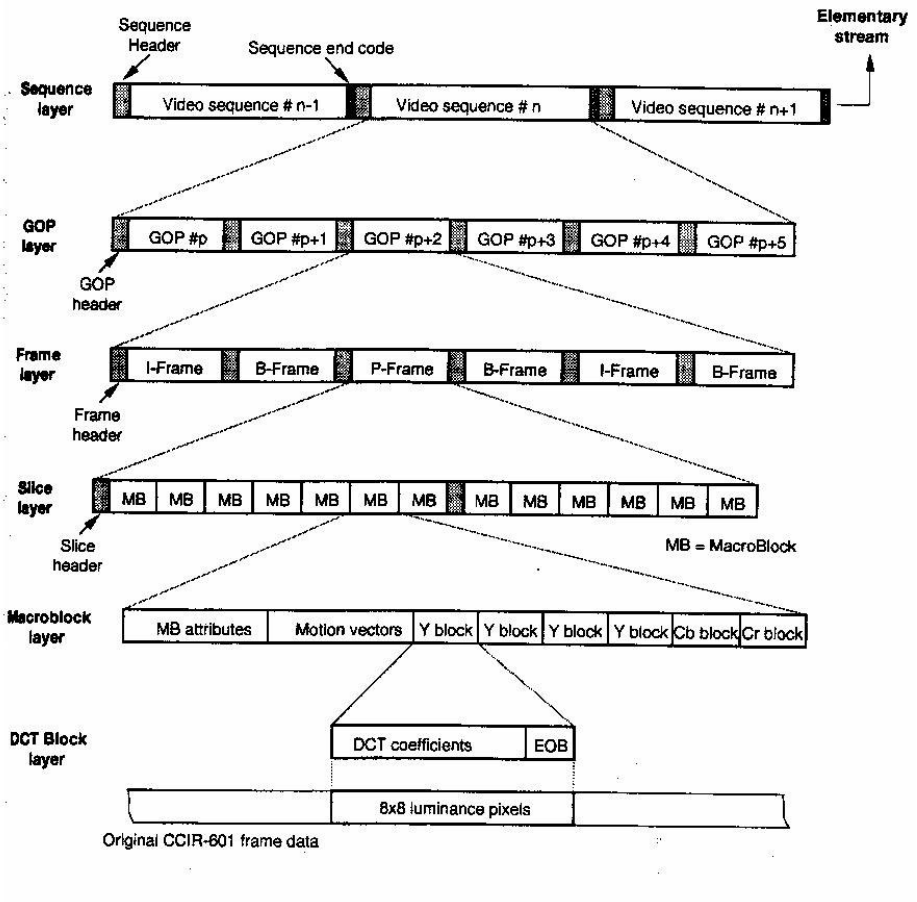


Figura 3.3

MPGE-4 [9] es la evolución de MPEG-1 y 2, y provee la tecnología base para la codificación en base a contenidos, y su almacenamiento, transmisión y manipulación. Presenta mejoras interesantes respecto a la eficiencia de la codificación, robustez de trasmisión e interactividad. MPGE-4 puede codificar múltiples “Objetos de video” (MVO – Multiple Video Objects), ya que sus contenidos son representados en forma individual. El receptor puede de esta manera recibir diferentes flujos por cada objeto codificado dentro de un mismo video, correspondientes por ejemplo a diferentes “planos” (VOP – Video Object Plane) de la imagen. Cada secuencia de VOPs constituye un objeto de video (VO – Video Object) independiente, los que son multiplexados dentro de una trasmisión, y demultiplexados y decodificados por el receptor.

En 2001, el grupo MPEG de ISO/IEC y el VCEG (Video Coding Expert Group) del ITU-T decidieron unir esfuerzos en un emprendimiento conjunto para estandarizar un nuevo codificador de video, mejor que los anteriores, especialmente para anchos de banda o capacidad de almacenamiento reducidos [10]. El grupo se llamó JVT (Joint Video Team), y culminó con la estandarización de la recomendación H.264/MPEG-4 Part 10, también conocida como JVT/H.26L/AVC (Advanced Video Coding) o H.264/AVC. Este nuevo estándar utiliza compensaciones de movimiento más flexibles, permitiendo dividir los macro-bloques en diversas áreas rectangulares, y utilizar desplazamientos

de hasta un cuarto de píxel. Agrega además los cuadros del tipo SP (Switching P) y SI (Switching I), similares a los P e I, pero con la posibilidad de reconstruir algunos valores específicos de forma exacta. Con AVC, para una misma calidad de video, se logran mejoras en el ancho de banda requerido de aproximadamente un 50% respecto estándares anteriores [11] [12].

En la Tabla 3.1 se presenta un resumen comparativo de los diferentes estándares de codificación de video. Como se puede observar en dicha tabla, los codificadores / decodificadores H.264/AVC no son compatibles con los estándares anteriores, lo que supone un punto de quiebre en la evolución del video digital.

Característica	**MPEG-1**	**MPEG-2**	**MPEG-4**	**H.264/MPEG-4 Part 10/AVC**
Tamaño del macro-bloque	16x16	16x16 (frame mode) 16x8 (field mode)	16x16	16x16
Tamaño del bloque	8x8	8x 8	16x16 8x8, 16x8	8x8, 16x8, 8x16, 16x16, 4x8, 8x4, 4x4
Transformada	DCT	DCT	DCT/DWT	4x4 Integer transfor
Tamaño de la muestra para aplicar la transformada	8x8	8x8	8x8	4x4
Codificación	VLC	VLC	VLC	VLC, CAVLC, CABAC
Estimación y compensación de movimiento	Si	Si	Si	Si, con hasta 16 MV
Perfiles	No	5 perfiles, varios niveles en cada perfil	8 perfiles, varios niveles en cada perfil	3 perfiles, varios niveles en cada perfil
Tipo de cuadros	I,P,B,D	I,P,B	I,P,B	I,P,B,SI,SP
Ancho de banda	Hasta 1.5 Mbps	2 a 15 Mbps	64 kbps a 2 Mbps	64 kbps a 150 Mbps
Complejidad del codificador	Baja	Media	Media	Alta
Compatibilidad con estándares previos	Si	Si	Si	No

Tabla 3.1

3.3. Transmisión de video sobre redes IP

3.3.1. Paquetización del video

Las secuencias (Elementary Streams) son paquetizadas en unidades llamadas PES (Packetized Elementary Streams), consistentes en un cabezal y hasta 8 kbytes de datos de secuencia. Estos PES a su vez, son paquetizados en pequeños paquetes, de 184 bytes, los que, junto a un cabezal de 4 bytes (totalizando 188 bytes) conforman el "MPEG Transport Stream" (MTS) y pueden ser transmitidos por diversos medios.

En redes IP, el transporte del video se realiza mediante los protocolos RTP y RTCP, descritos a continuación. El RFC 2250 [13] establece los procedimientos para transportar video MPEG-1 y MPEG-2 sobre RTP. Varios paquetes MTS de 188 bytes pueden ser transportados en un único paquete RTP, para mejorar la eficiencia.
Los RFC 3016 [14] y RFC 3640 [15] establecen los procedimientos para transportar flujos de audio y video MPEG-4.

3.3.2. RTP/RTCP – Real-Time Transport Protocol

El protocolo RTP, basado en el RFC 3550 [16], establece los principios de un protocolo de transporte sobre redes que no garantizan calidad de servicio para datos "de tiempo real", como por ejemplo voz y video.

El protocolo establece la manera de generar paquetes que incluyen, además de los propios datos de "tiempo real" a transmitir, números de secuencia, marcas de tiempo, y monitoreo de entrega. Las aplicaciones típicamente utilizan RTP sobre protocolos de red "no confiables", como UDP. Los datos obtenidos de cada conjunto de muestras de voz o video son encapsulados en paquetes RTP, y cada paquete RTP es a su vez encapsulado en segmentos UDP. Dentro del cabezal RTP se indica con 7 bits el tipo de información transportada (PT = Payload Type), lo que permite diferenciar hasta 128 tipos de información. Los valores de este campo se definen en el RFC 3551 [17]. Algunos valores de ejemplo se muestran en la Tabla 3.2

Payload Type	Formato	Medio	Clock Rate
0	PCM mu-law	Audio	8 kHz
3	GSM	Audio	8 kHz
4	G.723	Audio	8 kHz
8	PCM A-law	Audio	8 kHz
9	G.722	Audio	8 kHz
14	MPEG Audio	Audio	90 kHz
15	G.728	Audio	8 kHz
18	G.729	Audio	8 kHz
26	Motion JPEG	Video	90 kHz
31	H.261	Video	90 kHz
32	MPEG-1 o 2 Elementary Stream	Video	90 kHz
33	MPEG-1 o 2 Transport Stream	Video	90 kHz
34	H.263	Video	90 kHz

Tabla 3.2

El RFC 3550 establece, además del protocolo RTP, un protocolo de control, RTCP, encargado de enviar periódicamente paquetes de control entre los participantes de una sesión. El protocolo RTCP tiene las siguientes funciones principales:

- Proveer realimentación acerca de la calidad de los datos distribuidos (por ejemplo, de la calidad percibida). Esta realimentación permite adaptar dinámicamente la codificación, o tomar acciones tendientes a solucionar problemas cuando se detecta degradación en la calidad de la comunicación
- Transporte del CNAME (Canonical Name) de cada originador. Este identificador permite asociar varios flujos RTP con el mismo origen (por ejemplo, flujos de audio y video provenientes del mismo emisor)
- Adaptar dinámicamente la frecuencia de envío de paquetes de control RTCP de acuerdo al número de participantes en la sesión. Dado que los paquetes se deben intercambiar "todos contra todos", es posible saber cuantos participantes hay, y de esta manera calcular la frecuencia de envíos de esto paquetes.

3.3.3. Ancho de banda en IP para video

Como se ha visto, la codificación digital de video utiliza algoritmos de compresión, los que generan codificación de largo variable y flujos de ancho de banda también variables. Para una aplicación determinada, el ancho de banda requerido en una red IP dependerá del tipo de codificación

utilizada (MPEG-1, 2, 4), de la resolución (tamaño de los cuadros SD, CIF, QCIF, etc), del tipo de cuantización seleccionado y del movimiento y textura de la imagen. Al ancho de banda propio de la señal de video se le debe sumar la sobrecarga de los paquetes IP, UDP y RTP y para la LAN, de las tramas Ethernet, todo lo que puede estimarse en aproximadamente un 25% adicional. A diferencia de la codificación de audio, donde los anchos de banda pueden calcularse en forma exacta en base únicamente en el codec utilizado, la codificación de video es estadística, y depende de la imagen transmitida, por lo que los cálculos son también aproximados y estadísticos. A modo de ejemplo, en la Figura 3.4, tomada de [18], se muestra como varía el ancho de banda requerido, para diversos codecs, en función de la calidad de la imagen, para la secuencia de video "Tempete", en resolución CIF (352 x 288) a 15 Hz. Puede verse como el ancho de banda puede variar entre unos 64 kbps a 1 Mbps para H.264/AVC, dependiendo de la calidad (PSNR[1]), y como, para una misma calidad, MPEG-2 requiere de aproximadamente el doble de ancho de banda que H.264/AVC.

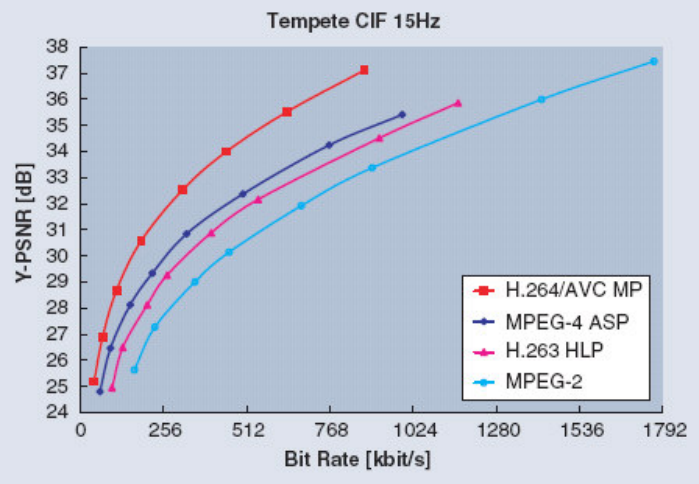


Figura 3.4

En general, el ancho de banda de las señales de video pueden variar desde valores cercanos a los 64 kbps (para baja resolución de pantalla, imágenes con poco movimento), hasta decenas de Mbps (para resoluciones medias o altas).

4. Calidad perceptual de video

La manera más confiable de medir la calidad de una imagen o un video es la evaluación subjetiva, realizada por un conjunto de personas que opinan acerca de su percepción. La opinión media, obtenida mediante el "MOS" (Mean Opinion Score) es la métrica generalmente aceptada como medida de la calidad. Para ello, los experimentos subjetivos controlados continúan siendo actualmente los métodos de medida reconocidos en la estimación perceptual de la calidad del video.

4.1. Métodos subjetivos de evaluación

Diversos métodos subjetivos de evaluación de video son reconocidos, y están estandarizados en las recomendaciones ITU-R BT.500-11 [19], especialmente desarrollada para aplicaciones de televisión y ITU-T P.910 [20], para aplicaciones multimedia. En todos los métodos propuestos, los

[1] La definición detallada de la métrica PSNR se presenta más adelante, en el capítulo 5

evaluadores son individuos que juzgan la calidad en base a su propia percepción y experiencia previa. Estos métodos tienen en común la dificultad y lo costoso de su implementación.

La recomendación BT.500-11 detalla los métodos DSIS, DSCQS, SSCQE y SDSCE. La P.910 los métodos ACR, DCR y PR. Todos ellos se describen brevemente a continuación.

Métodos propuestos en ITU-R BT.500-11

Escala de degradación con doble estímulo (DSIS – Double Stimulus Impairment Scale)

El método consiste en la comparación de dos estímulos, uno dado por la señal original (no degradada) y el otro por la señal degradada. En forma genérica, las señales pueden ser una imagen, una secuencia de imágenes o un video. Las condiciones de visualización, iluminación del ambiente, disposición de las personas respecto al monitor o televisor, etc. están controladas y detalladas en la recomendación.
Los participantes deben seleccionar una de entre cinco opciones, como se muestra en la Tabla 4.1:

La degradación es imperceptible	5
La degradación es perceptible, pero no molesta	4
La degradación es ligeramente molesta	3
La degradación es molesta	2
La degradación es muy molesta	1

Tabla 4.1

La estructura de la prueba consiste en mostrar la señal de referencia por unos segundos, y luego la señal degradada, por la misma cantidad de segundos. La misma secuencia se repite dos veces para cada prueba. Se solicita a los participantes que esperen hasta el final de cada secuencia de prueba para realizar su calificación (es decir, deben realizar la calificación luego de ver las dos secuencias de señal original / señal degradada).

Escala de calidad continua de doble estímulo (DSCQS – Double Stimulus Continuos Quality Scale)

Al igual que en el método anterior, se presentan dos señales. Sin embargo, en este método se solicita a los participantes que califiquen la calidad de ambas señales, en lugar de la degradación. Es decir, se debe calificar tanto la señal de referencia (señal "A") como la señal procesada o degradada (señal "B"). La calificación se realiza en base a una escala continua, utilizando una plantilla impresa de 10 cm de largo, donde se presentan las calificaciones indicadas en la Figura 4.1:

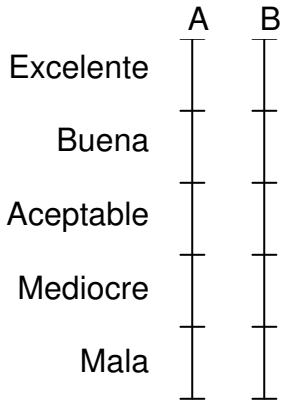


Figura 4.1

La escala continua permite medir diferencias más precisas entre ambas señales, e incluso permite categorizar a la segunda señal con mejor calidad que la primera. Esto último podría ser posible en el caso que se utilice este método para evaluar algoritmos de realce de video, o algoritmos que intenten mejorar la calidad (por ejemplo, compensar el "efecto de bloques", que se detallará más adelante).

Evaluación de calidad continua de estímulo único (SSCQE– Single Stimulus Continuous Quality Evaluation)

A diferencia de los métodos anteriores, en este método, se presenta una única secuencia de video a ser evaluada. Este video puede o no tener degradaciones.
También a diferencia de los otros métodos, se propone aquí una evaluación continua de la calidad del video, y no sólo una calificación global que integra las degradaciones de varios segundos. Para ello se utiliza un cursor móvil, conectado a una computadora, que permite registrar en forma continua las calificaciones.
Se utiliza la misma escala presentada en la
Figura 4.1. Dado que se toman muestras de la evaluación en forma continua, se puede asignar a cada instante del video su correspondiente calificación, lo que permite tener en forma mucho más detallada el efecto perceptual de cada una de las degradaciones.

Método de doble estímulo simultáneo para evaluación continua (SDSCE - Simultaneous Double Stimulus for Continuous Evaluation)

Cuando hay que evaluar la fidelidad, es necesario compara una señal contra su referencia. El SDSCE ha sido elaborado a partir del SSCQE, con ligeras diferencias en cuanto a la manera de presentar las imágenes a los sujetos y con respecto a la escala de apreciación. El método fue propuesto a MPEG para evaluar el comportamiento frente a errores a velocidades de transmisión muy bajas, pero puede ser aplicado adecuadamente a todos los casos en los que hay que evaluar la fidelidad de la información visual afectada por la degradación que varía en función del tiempo.
Con este método, el grupo de personas observa dos secuencias al mismo tiempo: una es la referencia, la otra es la señal degradada a evaluar. Ambas pueden ser presentadas dentro de un mismo monitor, o en dos monitores alineados. Se pide a los sujetos que comprueben las diferencias entre las dos secuencias y juzguen la fidelidad de la señal a calificar moviendo el cursor de un dispositivo de voto manual. Cuando la fidelidad es perfecta, el cursor debe estar en la parte superior de la escala (codificada con el valor100), cuando la fidelidad es nula, el cursor debe estar en la parte inferior de la escala (codificada con el valor 0).

Métodos propuestos en ITU-T P.910

Índices por categorías absolutas (ACR - Absolute Category Rating)

El método ACR es un juicio de categorías en el que las secuencias de prueba se presentan una por vez y se califican independientemente en una escala de categorías. Se utiliza una escala de 5 niveles, como se presenta en la Tabla 4.2

Excelente	5
Bueno	4
Aceptable	3
Mediocre	2
Mala	1

Tabla 4.2

Índices por categorías de degradación (DCR - Degradation Category Rating)

Es un método de doble estímulo, donde las secuencias se presentan por pares: el primer estímulo presentado en cada par es siempre la señal de referencia, mientras que el segundo estímulo es la señal degradada. Se pueden presentar las señales de referencia y la degradada en forma serial, una a continuación de la otra, o en forma conjunta, en el mismo monitor. La evaluación se realiza con la escala de 5 valores presentada en la Tabla 4.1

Método de comparación por pares (PC - Pair Comparison)

El método se utiliza cuando se desean comparar degradaciones producidas por dos sistemas diferentes, sobre una misma señal de referencia. En el método PC se presenta una señal luego de pasar por un sistema, y a continuación la misma señal luego de pasar por el otro sistema. Estos sistemas pueden ser simplemente codificadores, medios de transmisión, etc. Después de ver cada par de secuencias, se hace una apreciación sobre qué señal sufrió “menos degradaciones”, en el contexto del escenario de prueba.

4.2. Métodos objetivos de evaluación

Los métodos subjetivos, presentados en la sección anterior, son costosos, difíciles de realizar, e impracticables en aplicaciones de tiempo real.
Por esto se hace necesario el uso de métodos objetivos y automáticos, que puedan predecir con fiabilidad la calidad percibida, en base a medidas objetivas tomadas en algún punto del sistema.

No se han encontrado, al momento de escribir el presente trabajo, métodos objetivos de la medida perceptual de calidad de video estandarizados, que apliquen a todos los casos y con buenos resultados respecto a las medidas subjetivas, lo que es parte de la motivación del presente estudio. Sin embargo, existen varias propuestas de métricas de medida, con diversa complejidad y precisión de sus resultados, las que serán presentadas en el capítulo 5. Como introducción, se puede decir que las métricas basadas en analizar una imagen degradada y compararla píxel a píxel con su imagen de referencia (MSE, PSNR), no son suficientes para estimar la calidad ***percibida*** de la misma. El sistema visual humano es extremadamente complejo, y puede detectar fácilmente algún tipo de distorsión, mientras que puede pasar por alto otras, dependiendo de diversos factores. Estos factores pueden incluir el tipo de aplicación (TV, video conferencia, etc.), el lugar de la imagen en donde se produce la degradación (generalmente las degradaciones son menos visibles en regiones con muchos detalles o “actividad espacial”, o con gran movimiento, y son más visibles en imágenes estacionarias, o en fondos poco texturados). Incluso la calidad percibida puede depender del tipo de dispositivo utilizado y del tamaño del monitor [21].
Dado que el sistema de visión humano juega un rol fundamental, una introducción al mismo será presentada en 4.4

En forma genérica, los métodos objetivos de medida de calidad pueden clasificarse según la disponibilidad total, parcial o nula de la señal original, como se detalla a continuación.

Métodos con disponibilidad total de la señal original (FR - Full Reference)
Estos métodos se basan en la disponibilidad de la señal original, la que puede ser contrastada con la señal degradada, cuadro a cuadro. Esto presupone una severa restricción al uso práctico de este tipo de métodos, ya que en varias aplicaciones reales esto no es posible. Los métodos que utilizan métricas del tipo FR (Full Reference) pueden ser utilizados para categorizar en forma objetiva un sistema de transmisión, un codec, el efecto de un reducido ancho de banda, o de diversos factores que degraden una señal, en ambientes controlados. Sin embargo, no son

adecuados para aplicaciones de tiempo real (TV, video conferencias, etc.), ya que no es posible tener las señales originales.

Métodos con disponibilidad parcial de la señal original (RR - Reduced Reference)
Se trata de enviar, junto con el video codificado, algunos parámetros que caractericen a la señal, y que sirvan de referencia en el receptor para poder estimar la calidad percibida. Puede pensarse en la reserva de un pequeño ancho de banda (comparado con el del video) para el envío de este tipo de información adicional.

Métodos sin disponibilidad de la señal original - NR (No Reference)
Las personas no necesitan señales de referencia, ni información adicional para juzgar la calidad de una señal de video. Se basan en sus experiencias previas, y en las expectativas que tengan respecto al sistema. De igual manera, estos métodos intentan estimar la calidad percibida basándose únicamente en el análisis de la señal recibida. Son los métodos más complejos de implementar, pero no requieren de otra información que la propia señal de video.

El "Video Quality Experts Group" (VQEG) [22] está realizando un interesante y exhaustivo trabajo en el estudio y comparación de desempeño de métricas objetivas, separando el estudio por áreas de aplicación, de acuerdo al tipo de servicio: FR-TV (Full Reference TV), RRNR-TV (Reduced Reference / No Reference TV), HDTV (High Definition TV) y Multimedia. Un detalle de los avances realizados por éste grupo de trabajo se presenta más adelante, en el capítulo 5.2

4.3. Degradaciones en video digital

El proceso de digitalización de video utiliza técnicas que transforman una secuencia de píxeles al dominio de la frecuencia espacial (DCT), cuantificando valores, descartando eventualmente componentes de alta frecuencia, y haciendo uso de técnicas de predicción y compensación de movimientos. Esto genera "ruido de cuantificación", el que puede degradar la imagen original a niveles perceptibles. Es este "ruido de cuantificación" [23], el que genera las clásicas degradaciones que pueden verse en imágenes y videos con alta compresión, entre ellos, el conocido "efecto de bloques", que hace ver a la imagen como un conjunto de bloques pequeños (este efecto se verá con más detalles en los próximos párrafos)

Los algoritmos de compresión utilizados actualmente en la codificación digital de video introducen varios tipos de degradaciones, las que se pueden clasificar según sus características principales [24]. Esta clasificación es útil para poder comprender las causas de las degradaciones y el impacto que tiene en la calidad percibida.

Cuando el video digital es transmitido por algún medio de transmisión no confiable (por ejemplo, redes de paquetes), se suman, adicionalmente, degradaciones introducidas por las características propias del medio (por ejemplo, demoras o pérdidas de paquetes). La combinación de ambas degradaciones es la que finalmente percibe el usuario.

Efecto de bloques (blocking)
El efecto de bloques es, quizás, la más notoria de las degradaciones percibidas en video digital. Este efecto tiene su origen al dividir la imagen en bloques para realizar la transformada DCT. El efecto de bloques se presenta como discontinuidades en los bordes de bloques adyacentes al reconstruir la imagen. Dentro de un mismo cuadro, cuanto más "gruesa" sea la cuantificación realizada, más visible es el efecto de bloques. El umbral de cuantificación a partir del cual es percibido el efecto de bloque depende del tipo de imagen y del movimiento, por lo que no es posible definir un valor estándar e independiente de otros factores. Generalmente el efecto es

menos percibido en imágenes con movimiento, o en lugares de mucho o muy poco brillo. Los coeficientes de bajas frecuencias espaciales, y particularmente el coeficiente de DC de la transformada DCT, son los que determinan en mayor grado la visibilidad del efecto de bloques.
En cuadros predictivos, el efecto se puede dar entre macro-bloques, presentado discontinuidades entre sus bordes, o dentro de una macro-bloque, entre los cuatro bloques que lo componen. Como las compensaciones de movimiento, generalmente, proveen una buena predicción para los componentes de baja frecuencia, el error de predicción cuantificado se reduce a cero dentro de un macro-bloque, y si éste tiene un contenido uniforme, no se produce el efecto de bloques entre los cuatro bloques internos.

Diversos estudios se han realizado y existen varias propuestas de detección y corrección del efecto de bloques. Idealmente, estos procesos deberían mejorar el efecto de bloques, y a la vez mantener los bordes reales y la definición general de la imagen (es decir, no introducir borrosidad). Asimismo, es muy importante minimizar la capacidad de proceso necesario, ya que deben realizare en tiempo real.

La idea general es detectar discontinuidades en los valores de bajas frecuencias espaciales entre bloques adyacentes, que no se correspondan con cambios o bordes reales en la imagen. Se han propuesto técnicas basadas en métodos estadísticos, (asumiendo un modelo probabilístico de los coeficientes DCT) [25] [26], y utilizar las teorías de BPOCS (Block Projection Onto Convex Sets) [27]. Adicionalmente, técnicas que utilizan transformadas de Wavelets también han sido propuestas [28], con buenos resultados.

Las técnicas utilizadas para video pueden ser las mismas que las usadas para imágenes fijas, ya que el efecto se presenta en ambos casos. Sin embargo, en video puede ser utilizada la información temporal [29], para mejorar los algoritmos de detección y eliminación del efecto de bloques.

La detección y cuantificación del efecto de bloques en imágenes y videos es uno de los índices mayormente utilizado para estimar la calidad perceptual, como se verá en 5.4.

Efecto de imagen de base (basis image)
En los casos en los que uno de los coeficientes de la DCT es muy prominente respecto a los otros, y al utilizar cuantizaciones “gruesas”, es posible que quede como resultado un único coeficiente, que se traduce, al decodificar, como uno de los 63 posibles patrones de imágenes base de la DCT (ver Figura 3.1).
Este efecto, tiene, a su vez, efectos colaterales. Un bloque que ha sufrido el efecto de imagen base, seguramente no se corresponderá con sus bloques adyacentes, acentuando el efecto de bloques, y el efecto de mosaico, que se detalla más adelante.

Borrosidad o falta de definición (Blurring)
La falta de definición, se manifiesta como una pérdida de los detalles de la imagen. Si bien esto puede estar dado por imágenes tomadas fuera de foco, también puede ser un efecto introducido por el proceso de digitalización. En este caso, se da cuando se suprimen los coeficientes DCT de mayor orden, que son los que aportan los detalles finos dentro de sus bloques. Esta degradación también puede aportar al efecto de bloques y al de mosaico.

Varios métodos se han propuesto para estimar la definición (o borrosidad) general de una imagen, los que pueden aportar a métricas de la calidad percibida. En [30] se presenta un método sencillo, que califica con un porcentaje entre 0 y 100 la definición general de cada cuadro, basado en el análisis de los coeficientes DCT. Alternativamente se han propuesto métodos que utilizan una medida directa, basada en el análisis y clasificación de los bordes detectados en la imagen, y estimando en base a ellos la definición (o borrosidad) general [31].
.

Color bleeding (Corrimiento del color)
La falta de definición de la información de luminancia resulta en la difusión de los detalles en un área visible. El mismo concepto aplicado a la crominancia, produce "manchas" de colores sobre áreas de colores contrastantes. Al igual que el caso anterior, este efecto se debe a la supresión de los coeficientes de alta frecuencia de los componentes de crominancia. Dado que la crominancia es sub-muestreada, el efecto se propaga en todo el macro-bloque.

Se han propuesto técnicas para detectar y corregir el efecto de corrimiento de color. En [32] y [33] se presentan algoritmos de corrección de corrimiento de color, y se realiza una comparación con la métrica PSNR (ver 5) entre la imagen de referencia y la degradada, y entre la imagen de referencia y la degradada luego de aplicar la corrección. Se observa una leve mejora en esta métrica antes y después de aplicar el algoritmo de corrección. Sin embargo, es muy difícil deducir la mejora en la calidad percibida en forma objetiva.

Efecto escalera y Ringing
Cuando la imagen contiene bordes diagonales respecto a los ejes verticales u horizontales, se puede presentar el "efecto escalera". Dado que las imágenes base de DCT forman patrones horizontales y verticales, no se adecuan de la mejor manera a bordes diagonales, si se utilizan técnicas de compresión que eliminan la información de alta frecuencia espacial. Cuando la diagonal se distribuye entre varios bloques, se forma un patrón del tipo "escalera". Cuando las secciones adyacentes al borde tienen alto contraste, el efecto es especialmente notorio, y toma el nombre de "Ringing".

Patrones de mosaicos (Mosaic Patterns)
El efecto "mosaico" se presenta cuando parecen no coincidir los bordes de todos o gran parte de los bloques de una imagen. Este efecto, está muy relacionado al efecto de bloques.

Contornos falsos
El efecto de cuantización de los valores de luminancia de los píxeles lleva a que, en zonas de transiciones graduales, se generen falsos contornos, en los lugares de transición de un valor cuantizado a otro. Este efecto se nota si la cantidad de niveles de cuantización es insuficiente.
Algo similar sucede con la cuantización de los valores de DC y de baja frecuencia de los coeficientes de DCT, en éste tipo de imágenes.
El efecto se ve como saltos de luminancia o tonalidad, en lugares donde debería haber una transición gradual. Puede resultar especialmente visible en monitores o televisores grandes.

Detectar y corregir los contornos falsos no es sencillo, ya que, además de detectar los contornos, se debe decidir si corresponden a un borde real, o a un efecto del proceso de digitalización. Adicionalmente, al intentar remover los contornos falsos, no se deben introducir otras degradaciones, como borrosidad o disminución de la definición general de la imagen. Un método recientemente propuesto [34] presenta éstas características.

Bordes falsos
Este efecto se presenta como consecuencia de transportar el efecto de bloques hacia cuadros predictivos, con compensación de movimiento. Si se produce el efecto de bloques en una imagen tomada como referencia para próximos cuadros, estas discontinuidades producidas entre los bloques, pueden ser convertidas a bordes falsos dentro de bloques predictivos, debido a la estimación de movimiento. El origen de este efecto se esquematiza en la Figura 4.2. Allí se puede ver como los bordes de los bloques se trasladan en cuadros predictivos (B, P) a regiones que quedan dentro de los bloques, produciendo el efecto de bordes falsos.

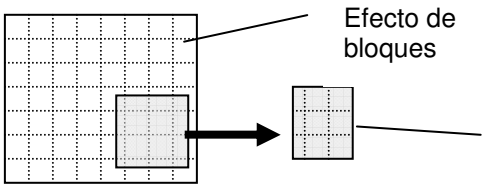


Figura 4.2

Errores de Compensaciones de Movimiento (MC mismatch)
La estimación de movimiento de un macro-bloque se realiza en el codificador, comparando el macro-bloque de una imagen con todos las posibles secciones de tamaño igual al macro-bloque (dentro de cierto rango espacial) de la(s) imagen(es) siguiente(s). La comparación se realiza generalmente buscando el mínimo valor de MSE (ver sección 5) entre el macro-bloque y la sección evaluada.

Este procedimiento (simple, pero de gran consumo de procesamiento), se basa en la hipótesis que todos los píxeles del macro-bloque tendrán un mismo desplazamiento, o sea, que corresponden a la misma figura en la imagen. Esto no es así cuando el macro-bloque contiene partes de diferentes figuras, o cuando contiene el borde entre una figura y un fondo fijo. En estos casos, la estimación de movimiento no será adecuada para una o quizás para ninguna de las figuras dentro del macro-bloque.

En estos casos, al reconstruir un macro-bloque basado en una mala estimación de movimiento, se puede ver un efecto de bloque, con componentes de alta frecuencia espacial, el que generalmente se aprecia sobre los bordes de figuras en movimiento.

Efecto mosquito
Se llama “efecto mosquito” a la fluctuación de la luminancia o crominancia alrededor de áreas de alto contraste o de figuras en movimiento. Este efecto está relacionado con los efectos de “ringnig” y de errores de compensación de movimiento, ya vistos. Es consecuencia de tener diferentes codificaciones en cuadros consecutivos, para una misma sección de la imagen (las que pueden darse por cambios en el tipo de predicción – hacia delante, hacia atrás, bidireccional - , niveles de cuantización, vectores de movimiento, etc.)

Fluctuaciones en áreas estacionarias
Fluctuaciones similares a las del efecto mosquito pueden verse también en áreas sin movimiento, pero con gran contenido de altas frecuencias espaciales (por ejemplo fondos con detalles pequeños, mucha textura, etc.)

De manera similar a lo explicado anteriormente, las fluctuaciones son consecuencia de los diferentes tipos de predicciones y niveles de cuantización utilizados entre cuadros. Estos efectos se ven enmascarados en áreas con movimiento, y por lo tanto solo se perciben en imágenes estáticas.

Errores de crominancia
Como se mencionó al explicar los errores de compensación de movimiento, la elección de los vectores de movimiento se basa en estimar el movimiento en base a minimizar los errores MSE de

la luminancia. Generalmente, la crominancia no es tenida en cuenta en esta estimación, aunque luego los valores estimados del movimiento son utilizados para los tres componentes de video. Esto puede llevar a que la estimación de movimiento no se adecue a la realidad, teniendo como consecuencia la aparición de macro-bloques de colores equivocados.

Resumen

Tal como fue descrito en los párrafos anteriores, el video puede presentar diversos tipos de degradaciones, generadas en el proceso de compresión y codificación, al utilizar técnicas que transforman una secuencia de píxeles al dominio de la frecuencia espacial (DCT), y cuantizar en forma "gruesa" los valores de los coeficientes. Cada tipo de degradación tiene un efecto perceptual diferente, y puede ser medido a su vez, de manera relativamente independiente. Varios métodos de medida del efecto de bloques y la pérdida de definición (borrosidad) se comparan en [35]. Según las conclusiones del citado artículo, ninguno de los métodos comparados presentan resultados satisfactorios con lo esperado en todos los aspectos evaluados, lo que demuestra la gran complejidad del problema de estimación de la calidad de video.

4.4. Sistema visual humano

El ojo humano es un sistema óptico formado por una dioptría esférica y una lente, que reciben, respectivamente, el nombre de córnea y cristalino, y son capaces de formar una imagen de los objetos sobre la superficie interna del ojo, en una zona denominada retina, que es sensible a la luz.

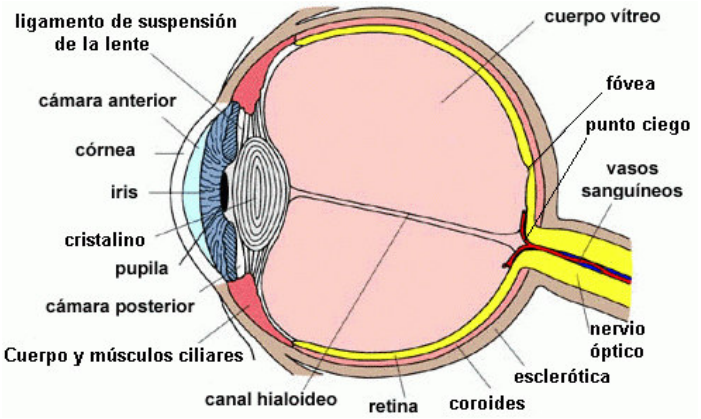


Figura 4.3

En la Figura 4.3 (tomada de [36]) se ven las partes que forman el ojo. Tiene forma aproximadamente esférica y está rodeado por una membrana llamada esclerótica que por la parte anterior se hace transparente para formar la córnea. Tras la córnea hay un diafragma, el iris, que posee una abertura, la pupila, por la que pasa la luz hacia el interior del ojo. El iris es el que define el color de nuestros ojos y el que controla automáticamente el diámetro de la pupila para regular la intensidad luminosa que recibe el ojo.

El cristalino está unido por ligamentos al músculo ciliar. De esta manera el ojo queda dividido en dos partes: la posterior que contiene humor vítreo y la anterior que contiene humor acuoso. El índice de refracción del cristalino es 1,437 y los del humor acuoso y humor vítreo son similares al del agua.

La córnea refracta los rayos luminosos y el cristalino actúa como ajuste para enfocar objetos situados a diferentes distancias. De esto se encargan los músculos ciliares que modifican la curvatura de la lente y cambian su potencia.

El cristalino enfoca las imágenes sobre la envoltura interna del ojo, la retina. Esta envoltura contiene fibras nerviosas (prolongaciones del nervio óptico) que terminan en unas pequeñas estructuras denominadas conos y bastones. Los bastones son detectores de luz muy sensibles. Las señales de muchos bastones pueden converger en una sola neurona, lo que mejora la sensibilidad pero reduce la resolución que se logra. Lo opuesto sucede con los conos. Varias neuronas transmiten las señales generadas por un único cono. Hay tres tipos de conos, los llamados "L", "M" y "S". Cada uno de estos tipos de conos tiene el pico de sensibilidad en diferentes partes del espectro, y son los responsables de la detección cromática, es decir, de los colores. Cada ojo tiene aproximadamente 5 millones de conos y 100 millones de bastones. Su densidad varia a lo largo de la retina. Los conos están concentrados en la fóvea, la zona central de la retina, mientras que los bastones dominan las áreas que rodean a la fóvea.

Los millones de nervios que van al cerebro se combinan para formar un nervio óptico que sale de la retina por un punto que no contiene células receptores. Es el llamado punto ciego.

El código neural que se genera por las células de la visión en la retina se transmite por vía del nervio óptico a los cuerpos geniculados y a la corteza visual en el cerebro. La corteza visual procesa y refina este código neural a información que determina el tamaño, contraste, forma, detalle, color, etc. El cerebro combina toda esta información para producir la percepción visual.

La respuesta del sistema visual humano depende en gran medida de las variaciones de la luminancia, mucho más que del valor absoluto de ésta. Esta característica se conoce como la *ley de Weber-Fechner*, que establece que *el menor cambio discernible en la magnitud de un estimulo es proporcional a la magnitud del estimulo*. En este caso, el contraste *C* es una medida de la variación relativa de la luminancia *L*, lo que se puede expresar como

$$C = \frac{\Delta L}{L} \qquad (4.1)$$

Los primeros modelos de visión humana, adoptaron una aproximación conocida como **"modelo mono canal"**. Esta teoría sostiene que el sistema visual humano puede representarse como un solo canal o filtro espacial, cuyas características están determinadas por la *función de sensibilidad de contraste* (CSF – Contrast Sensititiy Function). Esta función modela la variación de la sensibilidad del sistema visual humano en función de las frecuencias espaciales que tenga el estímulo visual observado. Estas frecuencias espaciales se miden en "ciclos/grado", entendiéndose que corresponden a variaciones de luminosidad en función del ángulo visual medido en grados. En el modelo mono canal, es aceptado que la función CSF tiene un comportamiento "pasa banda", como el indicado en la Figura 4.4 (tomada de [38]), con la mayor sensibilidad entre 1 y 10 ciclos/grado. Sin embargo, la forma exacta de esta curva depende fuertemente de diversos factores, entre ellos, de la iluminación, la orientación espacial de la imagen respecto a la retina, el tamaño de la imagen observada, y el tiempo por el que es observada la imagen. Adicionalmente, la función de sensibilidad también depende de los componentes cromáticos y no solo de la luminancia. En este caso, la forma de la curva se asemeja más a la de un filtro "pasa bajos". En aplicaciones de video es importante modelar la sensibilidad de contraste también con imágenes que varían en el tiempo. En la Figura 4.5, tomada de [38], se muestra un esquema de la CSF acromática y cromática en función de la frecuencia espacial y de la frecuencia temporal. Se puede ver como en el caso acromático, para frecuencias temporales muy bajas (imágenes fijas), la gráfica se aproxima a una función pasa banda, similar a la de la Figura 4.4, mientras que en el caso cromático, para frecuencias temporales muy bajas, la gráfica se aproxima a un filtro pasa bajos.

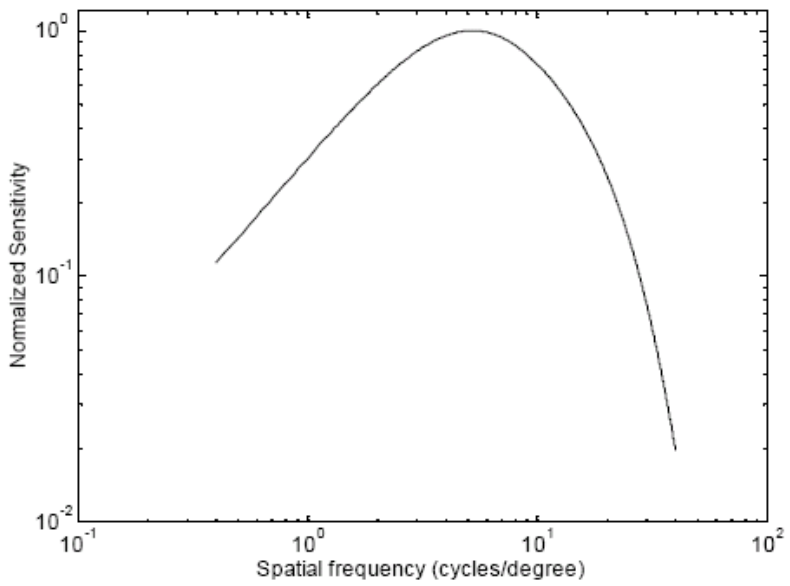
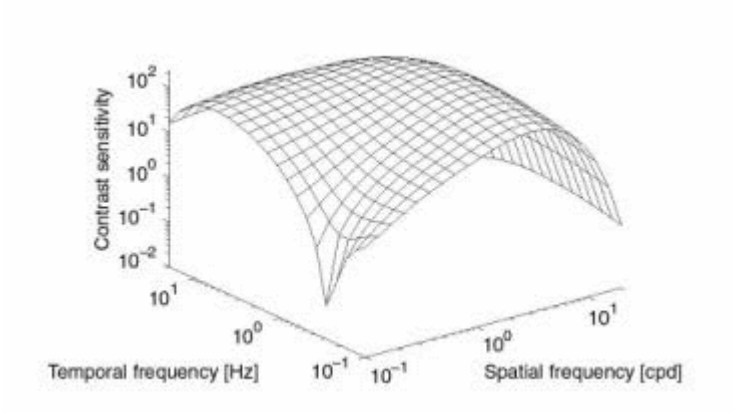
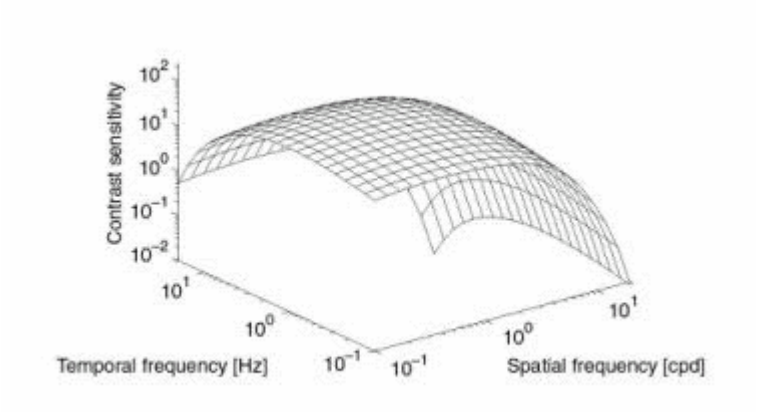


Figura 4.4



a) CSF acromática



b) CSF cromática

Figura 4.5

Una nueva teoría del modelo de visión humana, conocida como **"modelo multi-canal"** sostiene que el sistema visual humano no dispone de un único filtro de frecuencia espacial, sino que en realidad existen varios filtros, centrados en diferentes frecuencias espaciales, y que trabajan en paralelo y en forma independiente. Según este modelo, antes de ser analizada, la imagen que capta la retina es descompuesta por un banco de filtros casi lineales, que producen como salida un conjunto de "imágenes" de banda limitada en frecuencia espacial. Estas imágenes filtradas son analizadas separadamente en el sistema de visión humano y la información extraída de cada una de ellas es utilizada directamente, o combinada con la de las demás, para facilitar una descripción útil del mundo. En este modelo, cada canal sólo transmite los componentes de la imagen que se hallan en una determinada banda de frecuencia espacial, mientras que atenúa los situados en otras bandas (los que a su vez son transmitidos por otros canales). Según este modelo, la función de sensibilidad de contraste CSF es esencialmente la envolvente de las funciones de sensibilidad de cada uno de los filtros.

5. Medida de la calidad perceptual de video

5.1. Introducción

Las primeras medidas objetivas de la calidad del video están basadas en obtener las diferencias, píxel a píxel, entre las imágenes originales (previo a la compresión y transmisión) y las imágenes presentadas (luego de la recepción y reconstrucción). Dado que los sistemas de video utilizan técnicas de compresión con pérdida de información, y que los medios de transmisión a su vez pueden introducir factores distorsionantes (retardos, pérdida de paquetes, etc.), las imágenes presentadas serán diferentes a las originales.

Las medidas más simples son las de error cuadrático medio (MSE - Mean Square Error) y su raíz cuadrada (RMSE = Root Mean Square Error) y la relación señal a ruido de pico (PSNR – Peak Signal to Noise Ratio), definidas en las ecuaciones (5.1) a (5.3) más adelante. Estas métricas son del tipo FR (Full Reference), ya que requieren de la referencia completa para poder ser calculadas.

$$MSE=\frac{1}{TMN}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{t=1}^{T}\left[x(m,n,t)\ -y(m,n,t\ )\right]^2 \qquad (5.1)$$

$$RMSE=\sqrt{MSE} \qquad (5.2)$$

$$PSNR=10\log_{10}\left(\frac{L^2}{MSE}\right) \qquad (5.3)$$

donde la imagen tiene N x M píxeles y T cuadros, x, y son los píxeles de la imagen original y la distorsionada respectivamente. L es el rango dinámico que pueden tomar los valores de x o y , y toma el valor 255 para 8 bits por píxel.

Estas métricas son fáciles de calcular, y tienen un claro significado. Por estas razones, han sido ampliamente usadas como métricas en la estimación de la calidad de video. Hay que poner especial énfasis en la alineación espacial y temporal de las imágenes a comparar, ya que la referencia y la imagen degradada pueden estar desfasadas en el tiempo o en el espacio.

Sin embargo, también han sido ampliamente criticadas por no tener correlación directa con la calidad percibida. Por ejemplo, en la Figura 5.1, tomada de [37], se muestran tres ejemplos de imágenes comprimidas, donde se puede ver claramente que con similares valores de MSE, la calidad percibida puede ser esencialmente diferente (comparar, por ejemplo, "Tiffany" con "Mandril", sobre el lado derecho de la figura), lo que pone en duda la utilidad de este tipo de métrica como indicador de calidad. En la Figura 5.2, presentada en [38], se puede ver como la misma imagen, con el mismo valor de PSNR, puede tener diferente calidad percibida, dependiendo del lugar en el que se presenten las degradaciones. En la figura (b), se nota claramente la degradación en el cielo (parte superior), mientras que en la figura (c), una degradación similar en la parte inferior prácticamente no es perceptible. Este fenómeno se conoce como "enmascaramiento". En zonas texturadas o con gran "actividad espacial", las degradaciones quedan "enmascaradas" y son menos percibidas por el sistema visual humano. El enmascaramiento también puede darse en video, donde cambios rápidos temporales pueden enmascarar cierta pérdida de calidad en cada cuadro.

Debido a la baja correlación entre el MSE, RMSE y PSNR con la calidad percibida, en los últimos tiempos, se ha realizado un gran esfuerzo para desarrollar nuevos modelos que tengan en cuenta las características de percepción del sistema visual humano y que permitan calcular métricas objetivas que lo simulen. La gran mayoría de estos modelos propuestos hasta el momento son del tipo FR (Full Reference), ya que requieren de la señal de referencia para el cálculo de sus métricas.

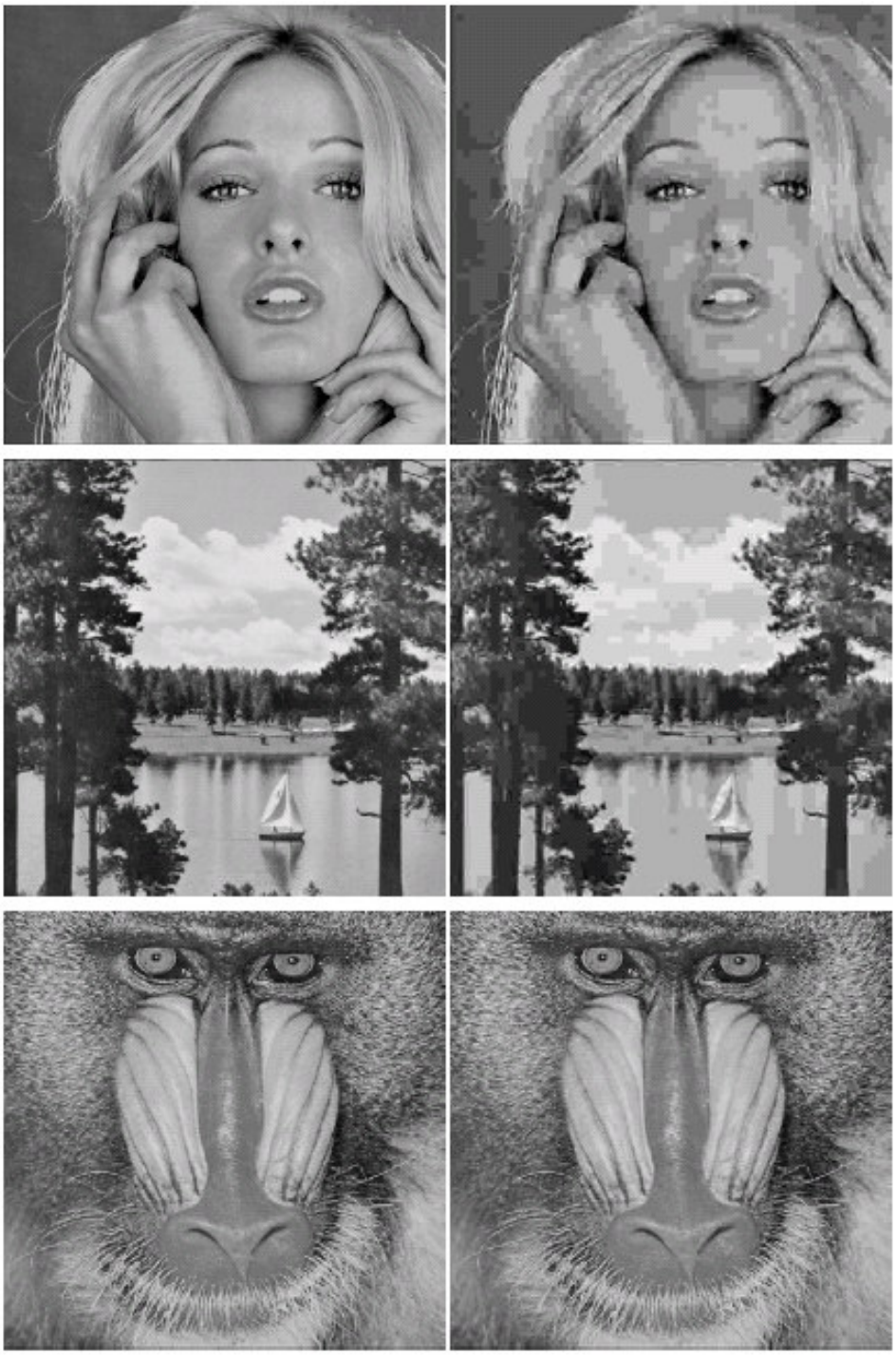


Figura 5.1
Evaluaciones de imágenes comprimidas con JPEG.
Arriba: Imagen “Tiffany” original y comprimida, MSE=165; Medio: Imagen “Lago” original y comprimida, MSE=167; Abajo: Imagen “Mandril” original y comprimida, MSE=163

Figura 5.2

5.2. El trabajo del VQEG

Varios aspectos han de ser evaluados al momento de decidir la utilización de un modelo. Una es la capacidad efectiva del modelo de predecir en forma consistente la calidad percibida, para diversos tipos de videos, ya sea con poco o mucho movimiento, con imágenes naturales o animadas, y bajo diferentes niveles de compresión. Otros factores a tener en cuenta son la capacidad de procesamiento, memoria y recursos requeridos.

El VQEG (Video Quality Expert Group) [22] está llevando a cabo un gran trabajo sistemático y objetivo de comparación de modelos. El objetivo del VQEG es proporcionar evidencia para los organismos internacionales de estandarización acerca del desempeño de diversos modelos propuestos, a los efectos de definir una métrica estándar y objetiva de calidad percibida de video digital (VQM – Video Quality Metric).

Los estudios del VQEG se dividen en

- FR-TV (Full Reference TV)
- RRNR-TV (Reduced Reference, No Reference TV)
- Multimedia
- HDTV

Cada aplicación, por sus propias características, requiere de pruebas y modelos diferentes. El primer proyecto, ya terminado, es el correspondiente a FR—TV. Este proyecto fue llevado a cabo en dos fases, como se detalla en la siguiente sección.

5.2.1. FR-TV (Full Reference TV)

En la fase I de FR-TV, llevada a cabo entre 1997 y 2000, se evaluaron 9 propuestas, además de la PSNR:

- Peak signal-to-noise ratio (PSNR, P0)
- Centro de Pesquisa e Desenvolvimento (CPqD, Brazil, P1, August 1998)
- Tektronix/Sarnoff (USA, P2, August 1998)
- NHK/Mitsubishi Electric Corporation (Japan, P3, August 1998)
- KDD (Japan, P4, model version 2.0 August 1998)
- Ecole Polytechnique Féderal Lausanne (EPFL, Switzerland, P5, August 1998)
- TAPESTRIES (Europe, P6, August 1998)
- National Aeronautics and Space Administration (NASA, USA, P7, August 1998)
- Royal PTT Netherlands/Swisscom CT (KPN/Swisscom CT, The Netherlands, P8, August 1998)
- National Telecommunications and Information Administration (NTIA, USA, P9, model version 1.0 August 1998)

Las evaluaciones fueron realizadas con diversos tipos de material de video, incluyendo 20 tipos diferentes de contenidos (entre los que hay deportes, animaciones, escenas de interiores y exteriores, etc.) y con velocidades de 768 kb/s hasta 50 Mb/s. Se utilizó el método DSCQS (Ver 4.1), y las pruebas se realizaron sobre una base de más de 26.000 opiniones subjetivas, en 8 laboratorios independientes en diferentes partes del mundo.

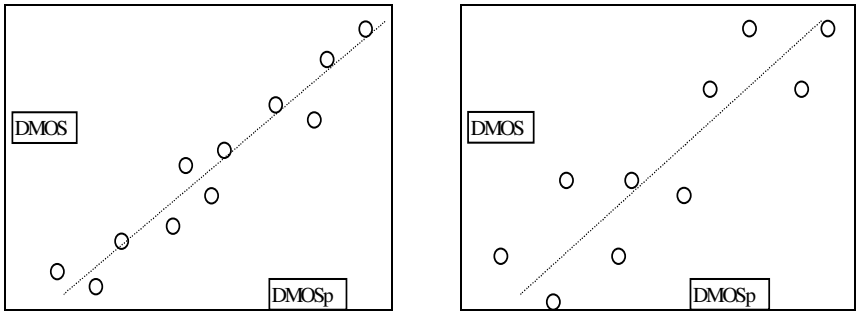
Como se mencionó, el objetivo es contrastar el resultado presentado por cada uno de los modelos propuestos, contra el resultado subjetivo obtenido para el mismo video, utilizando el método DSCQS. El desempeño de cada una de los modelos propuestos fue evaluado respecto a tres aspectos de su capacidad de estimar la calidad subjetiva:

- **Precisión** de la predicción: La capacidad de predecir la calidad subjetiva con mínimo error. Se puede determinar numéricamente mediante el coeficiente de correlación lineal de Pearson, definido a continuación

$$r_p = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \qquad (5.4)$$

  donde $\bar{x}$ y $\bar{y}$ son los valores medios de cada juego de datos (por ejemplo x puede representar los datos obtenidos como resultado del modelo a evaluar y y los datos obtenidos mediante las pruebas subjetivas). Si la relación entre los valores de x, y es lineal, r_p toma el valor 1.

  A modo de ejemplo, en la Figura 5.3, se presentan dos posibles gráficas, donde se muestra el valor obtenido mediante la prueba subjetiva (DMOS) y el valor obtenido por el modelo (DMOSp). Idealmente, deberían estar todos los puntos sobre una recta a 45 grados. En el ejemplo, el modelo representado en la figura (a) se desempeña mejor que el de la figura (b), ya que los puntos de la gráfica (a) se aproximan mejor a una recta a 45 grados.

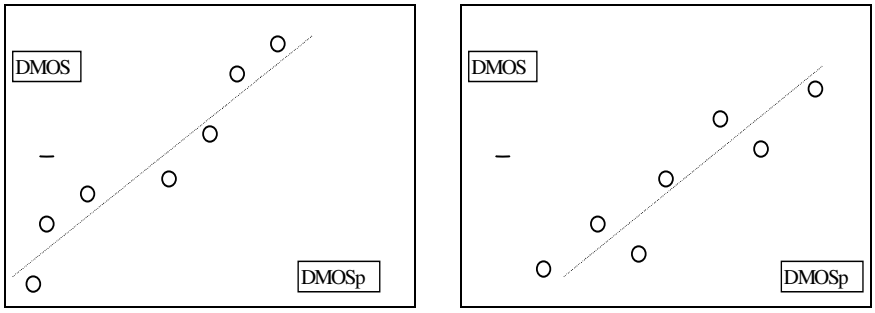


(a) (b)

Figura 5.3

- **Monotonicidad**: Mide si los incrementos (decrementos) en una variable están asociados a incrementos (decrementos) en la otra. Típicamente se mide con el coeficiente de correlación de Spearman.

  En la Figura 5.4 se presentan dos posibles gráficas, donde se muestran valores hipotéticos obtenidos mediante la prueba subjetiva (DMOS) y valores obtenidos por el modelo (DMOSp). Ambas tiene similar precisión, pero el modelo representado en la figura (a) presenta monotonicidad, mientras que el de la figura (b) no.



(a) (b)

Figura 5.4

- **Consistencia**: Se corresponde con el grado en que el modelo mantiene la precisión a lo largo de las secuencias de pruebas. Se puede evaluar midiendo la cantidad de puntos para los que el error de la predicción es mayor a cierto umbral, por ejemplo, el doble de la desviación estándar, como se indica a continuación:

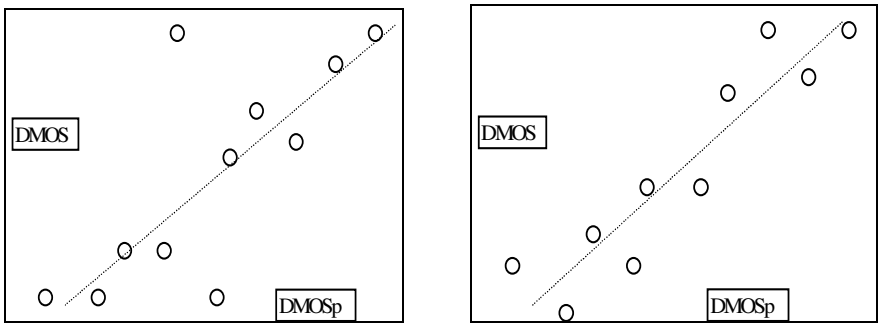
$$\left|x_i - y_i\right| > 2\sigma_{y_i} \qquad (5.5)$$

  El grado de consistencia se puede definir entonces, como el cociente entre los puntos alejados más de 2σ (N_o) respecto al total de puntos (N). Evidentemente, cuanto más próximo a 0 sea este coeficiente, mejor resultará el modelo:

$$r_0 = \frac{N_0}{N} \qquad (5.6)$$

En la Figura 5.5 se presentan dos posibles gráficas, donde se muestra el valor obtenido mediante la prueba subjetiva (DMOS) y el valor obtenido por el modelo (DMOSp). En el modelo representado en la figura (a) se puede ver que la mayoría de los puntos se encuentran "razonablemente" cerca de la recta, mientras que dos puntos (sobre el centro de la gráfica) están muy alejados. En la figura (b) todos los puntos se encuentran a distancias similares de la recta, sin aparecer puntos especialmente alejados. Ambas distribuciones pueden tener iguales valores de RMS, pero el modelo (b) tendrá mayor "consistencia", ya que los errores se distribuyen de manera más uniforme que en el modelo de la figura (a), donde la predicción es mejor que el modelo (b) para la mayoría de los puntos, pero mucho peor en 2 de los puntos.

Otra métrica utilizada para medir la consistencia es el valor de Kurtosis, que es una cantidad adimensionada relacionada con la forma de la gráfica de la distribución de errores, en forma independiente del valor cuadrático medio del error. Valores altos de Kurtosis indican que existen desviaciones extremas infrecuentes en la distribución, mientras que valores bajos indican desviaciones más frecuentes pero de valores más pequeños.



(a) (b)

Figura 5.5

Como resultado de la fase I de FR-TV, dependiendo de la métrica de comparación utilizada, siete u ocho de los modelos propuestos resultaron estadísticamente equivalentes entre sí, y a su vez, equivalentes a los resultados obtenidos con el PSNR [39]. Este resultado fue realmente desalentador, ya que indica que no existen diferencias apreciables entre el sencillo cálculo del PSNR y los sofisticados métodos preceptuales propuestos. En base a estos resultados, el VQEG ha realizado una segunda fase de pruebas, llamando nuevamente a interesados en contrastar sus modelos. La denominada "fase II" para FR-TV fue realizada entre los años 2001 y 2003 y los resultados finales fueron publicados en agosto de 2003 [40]. El objetivo de esta segunda fase era obtener resultados más discriminatorios que los obtenidos en la fase I. Al igual que en la fase I, las evaluaciones fueron realizadas con diversos tipos de material de video, y en formato de 525 y 625 líneas por cuadro. Se evaluaron velocidades entre 768 kb/s y 5 Mb/s. Las pruebas fueron realizadas en 3 laboratorios independientes, en Canadá, Estados Unidos e Italia.

Se evaluaron 6 proponentes, algunos de los cuales ya habían sido evaluados en la fase I:

- NASA (USA, Proponent A)
- British Telecom (UK, Proponent D)
- Yonsei University / Radio Research Laboratory (Korea, Proponent E)
- CPqD (Brazil, Proponent F)
- Chiba University (Japan, Proponent G)
- NTIA (USA, Proponent H)

En la Figura 5.6 se muestra la exactitud de las predicciones de cada uno de estos modelos, medidos con el coeficiente de correlación de Pearson, según los datos presentados en [40]. Puede verse que, en el formato de 525 líneas, los modelos de British Telecom y de NTIA se destacan del resto, y son estadísticamente equivalentes, según las conclusiones del informe del VQEG. Estos modelos presentan una mejora de un 17% respecto al PSNR. Por otra parte, en el formato de 625 líneas, los modelos de NASA, Yonsei, CpQD y NTIA son los mejores, y también son estadísticamente equivalentes, según las conclusiones del informe. En promedio, presentan una mejora de un 21% respecto al PSNR. Con otras métricas de comparación de los modelos (monotonicidad, consistencia) se obtienen resultados similares.

Es de hacer notar que no se evidencia, a priori, ninguna razón objetiva por la que la calidad percibida deba ser diferente según la cantidad de líneas de la imagen. Algunos modelos han presentado mejores resultados para 525 líneas, mientras que otros lo han hecho para 625 líneas.

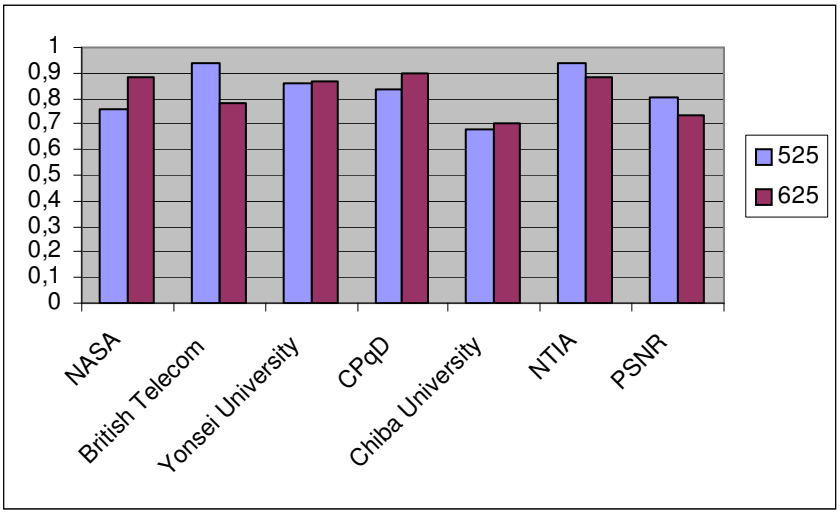


Figura 5.6

Sobre la base de estos resultados, la ITU ha estandarizado, en la recomendación ITU-R BT.1683 [41] de junio de 2004, a los cuatro mejores algoritmos, definidos en el marco de esta recomendación como:

- British Telecom BTFR (Reino Unido)
- Yonsei University/Radio Research Laboratory/SK Telecom (Corea)
- Centro de Investigación y Desarrollo en Telecomunicaciones (CPqD) (Brasil)
- National Telecommunications and Information Administration (NTIA)/Institute for Telecommunication Sciences (ITS) (Estados Unidos)

Es importante destacar que los modelos propuestos en esta Recomendación pueden utilizarse para evaluar un códec (combinación de codificador/decodificador) o una combinación de varios métodos de compresión y dispositivos de almacenamiento en memoria. Aunque en el cálculo de los estimadores de la calidad objetiva descrito en la Recomendación se considera la degradación provocada por errores (por ejemplo, errores en los bits, paquetes rechazados), no se dispone aún de resultados de pruebas independientes para validar la utilización de estimadores en los sistemas con degradación por errores. El material de pruebas utilizado por el VEQG no contenía errores de canal.

A continuación se presenta una breve descripción de los modelos incluidos en la recomendación ITU-R BT.1683:

British Telecom BTFR

El algoritmo BTFR (British Telecom Full Reference) consiste en una detección seguida de una integración. La detección supone el cálculo de un conjunto de parámetros perceptuales significativos del detector, a partir de secuencias de vídeo sin distorsionar (de referencia) y distorsionadas (degradadas). Estos parámetros se introducen a continuación en el integrador que produce una estimación de la calidad de vídeo percibida mediante la ponderación de cada uno de los parámetros. La selección de detectores y factores de ponderación se basa en el conocimiento de las propiedades de enmascaramiento espacial y temporal del sistema de visión humano. El proceso se esquematiza en la Figura 5.7

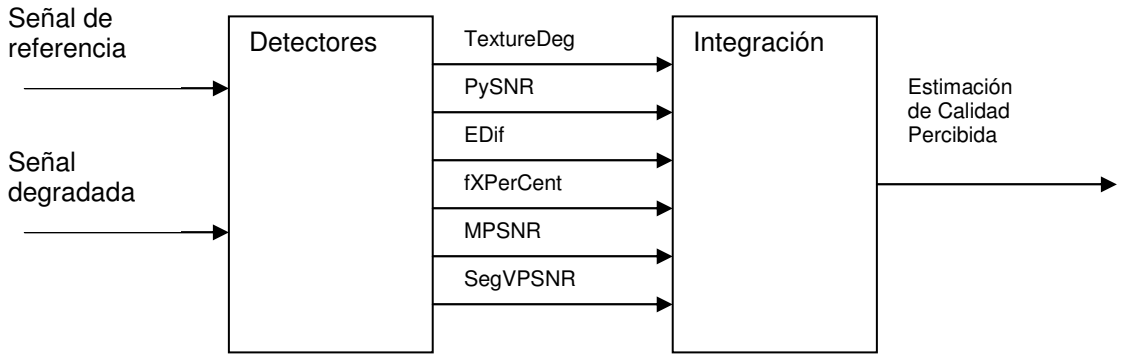


Figura 5.7

En la etapa de "Detectores" se incluye un proceso de recorte y desplazamiento, en el que básicamente las señales de referencia y degradada se alinean espacialmente. Esto es seguido de una etapa de emparejamiento. El proceso de emparejamiento genera señales para otros procedimientos de detección así como parámetros de detección para el procedimiento de integración. Las señales emparejadas se generan a partir de un proceso de búsqueda del emparejamiento óptimo para bloques de 9 x 9 píxeles de cada campo degradado. Este proceso genera una secuencia, denominada referencia emparejada, que sustituye a la secuencia de referencia en algunos módulos de detección.

TextureDeg (Análisis de textura): La textura de las secuencias se mide registrando el número de puntos de inflexión de la señal de intensidad a lo largo de las líneas horizontales de imagen. Una vez procesadas todas las líneas de un campo, un contador contiene el número de puntos de inflexión de la señal de intensidad horizontal, que se utiliza para calcular el parámetro de textura para cada campo, normalizándolo de acuerdo a la cantidad de píxeles horizontales y líneas verticales.

PySNR (SNR Piramidal): Se corresponde con un análisis espacial de frecuencia. Un detector espacial de frecuencia se basa en una transformación "piramidal" de las secuencias de referencia degradada y emparejada. En primer lugar se transforma cada secuencia generando las matrices piramidales de referencia y degradada. A continuación se calculan las diferencias entre las matrices piramidales midiendo los errores cuadráticos medios, obteniendo los resultados como SNR piramidal.

Edif (Análisis de bordes): Cada uno de los campos de las secuencias de referencia degradada y emparejada se somete por separado a una rutina de detección de bordes, utilizando el detector propuesto por Canny [42], para generar los correspondientes mapas de bordes y compararlos entre sí. Estos mapas de bordes son matrices con el valor 1 en el punto (x,y) donde se ha detectado un borde, 0 en todos los otros puntos. A continuación se calcula un estimador de la medida de las diferencias de bordes en todo el campo.

FXPerCent (Estadísticas de Emparejamiento): Se elaboran estadísticas de emparejamiento horizontal a partir del proceso de emparejamiento para utilizarlas en el proceso de integración. El mejor emparejamiento para cada bloque de análisis se utiliza en la construcción de un histograma de emparejamiento. Para cada campo, la medición *fXPerCent* es el porcentaje de bloques emparejados que contribuye al pico del histograma de emparejamiento.

MPSNR y SegVPSNR: Corresponden al cálculo del PSNR de la señal emparejada y de la PSNR segmentaria del campo, respectivamente

Una vez calculados los 6 factores detallados anteriormente, se aplican a la etapa de integración, la que implementa una suma ponderada de un promedio temporal de cada uno de ellos, según la siguiente ecuación

$$PDMOS=Offest+\sum_{k=0}^{5} AvD(k).W(k) \qquad (5.7)$$

donde $AvD(k)$ es el promedio del parámetro k (TextureDeg, MPSNR, etc) y $W(k)$ es un factor de ponderación, que, al igual que el Óffest, esta predeterminado en la recomendación según el tipo de video (525 o 625 líneas)

Yonsei University/Radio Research Laboratory/SK Telecom

El modelo propuesto por Yonsi University se basa en la observación que el sistema visual humano es altamente sensible a la degradación en torno a los bordes. Dicho de otro modo, cuando las zonas de los bordes de un video o imagen son borrosas, los evaluadores tienden a otorgar puntaciones menores al video aunque el error cuadrático medio global sea pequeño. Se observa además que los algoritmos de compresión de video tienden a producir mayores degradaciones en torno a las zonas de los bordes (ver 4.3, donde se puede ver que varios de los tipos de degradaciones se ven magnificadas en los bordes). De acuerdo con esta observación, el modelo proporciona un método de medición objetiva de la calidad de vídeo que **mide la degradación en torno a los bordes**. En este modelo, se aplica en primer lugar un algoritmo de detección de bordes a la secuencia de vídeo fuente para localizar las zonas de los bordes. A continuación, se mide la degradación de dichas zonas de bordes calculando el error cuadrático medio. A partir de dicho error se calcula la EPSNR del borde, que se utiliza como métrica de la calidad de vídeo tras una etapa de post-procesamiento.

En este modelo, es necesario aplicar primero un algoritmo de detección de bordes para localizar las zonas de los bordes. Se puede utilizar cualquier algoritmo de detección de bordes, aunque puede haber diferencias de menor importancia en los resultados. Por ejemplo, se puede utilizar cualquier operador gradiente para localizar las zonas de los bordes. Los píxeles cuyos gradientes superen en magnitud un cierto valor umbral se consideran zonas de bordes. Con estos píxeles puede aplicarse una “máscara”, con valor 1 en los píxeles cuyo gradiente supera el umbral, y con valor 0 en todos los otros puntos.

A continuación, se calculan las diferencias entre la secuencia de vídeo fuente y la secuencia de vídeo procesada, correspondiente a los píxeles no nulos de la máscara. Dicho de otro modo, se calcula el error cuadrático de las zonas de bordes del siguiente modo:

$$se_e^l = \sum_{i=1}^{M}\sum_{j=1}^{N}\left[S^l(i,j) - P^l(i,j)\right]^2 \qquad \text{si} \qquad \left|R^l(i,j)\right| \neq 0 \tag{5.8}$$

$$mse_e = \frac{1}{K}\sum_{l=1}^{L} se_e^l \tag{5.9}$$

$$EPSNR = 10\log_{10}\left(\frac{P^2}{mse_e}\right) \tag{5.10}$$

donde
$S^l(i, j)$: l-ésima imagen de la secuencia de vídeo fuente
$P^l(i, j)$: l-ésima imagen de la secuencia de vídeo procesada
$R^l(i, j)$: l-ésima imagen de la secuencia de vídeo máscara
M : número de filas
N : número de columnas.
L : número de imágenes (cuadros o campos)
K : número total de píxeles de las zonas de bordes
P: valor pico de píxel (255)

El valor obtenido del EPSNR es luego pasado por una etapa de post-proceso, donde se aplica un "de-énfasis" o un ajuste que toma en consideración bordes borrosos.

El modelo es extremadamente rápido. Una vez generado el mapa de bits, el modelo es varias veces más rápido que la PSNR convencional, lo que supone una mejora adicional importante.

CpQD-IES (Image Evaluation based on Segmentation)

Este modelo se basa en la segmentación de las escenas naturales en regiones planas, de bordes y de textura [43], y se asigna un conjunto de parámetros objetivos a cada uno de estos contextos. El resultado final se obtiene a partir de una combinación lineal de los niveles de degradación estimados, siendo el peso de cada nivel de degradación proporcional a su fiabilidad estadística.

Inicialmente se realiza una corrección de desplazamiento espacial y temporal y un ajuste de ganancia en la señal degradada.

La primera etapa del proceso es la segmentación de la imagen. Inicialmente, el algoritmo de segmentación clasifica cada píxel de la componente *Y* de un cuadro determinado de la escena original en regiones planas y no planas. Además, el algoritmo aplica a *Y* un detector de bordes, quedando definida la región de bordes dentro de los límites de la región plana. La región de textura está compuesta por los restantes píxeles de la imagen *Y.*

Se calcula luego la imagen de la magnitud del gradiente de Sobel calculado para una componente determinada (Y, Cb o Cr) de un cuadro determinado de la escena original, y la imagen de la magnitud del gradiente de Sobel de la misma componente del cuadro de la escena degradada. Luego se calcula la diferencia absoluta entre las imágenes de gradiente dentro de cada región de píxeles que pertenece a un contexto determinado (plano, borde o textura). Se define la diferencia

absoluta de Sobel (ASD - Absolute Sobel Difference) para esta componente de imagen y contexto como el promedio de las diferencias, restringidas a cada contexto (plano, borde o textura).
De este procedimiento se obtiene un conjunto de nueve mediciones objetivas {m1, m2, ..., m9} para cada cuadro de imagen, considerando los tres contextos y las tres componentes de imagen (Y, Cb y Cr).

El sistema CPqD-IES utiliza una base de datos de modelos de degradación a fin de estimar el índice de calidad de vídeo de la imagen degradada. Esta base de datos está integrada por información relativa a 12 escenas que representan diversos grados de movimiento (escenas dinámicas y estáticas), naturaleza (escenas reales y sintéticas), y contexto (cantidad de píxeles de textura, plano y bordes). Estas imágenes están "calibradas" con medidas subjetivas, y sirven como referencia para el cálculo del estimador de la calidad percibida para cualquier imagen.

NTIA - National Telecommunications and Information Administration

Sobre la base de los resultados del VQEG, el modelo propuesto por la NTIA [44] fue estandarizado por el American National Standards Institute (ANSI) en 2003. Este modelo tiene parámetros objetivos para medir los efectos perceptuales de una amplia gama de degradaciones tales como la borrosidad, el efecto de bloques, el movimiento entrecortado o poco natural, el ruido (tanto en la luminancia como en la crominancia), y los bloques con errores (por ejemplo, lo que podría verse normalmente cuando hay errores en la transmisión digital). Este modelo consta de una combinación lineal de siete parámetros. Cuatro de ellos se basan en características extraídas de los gradientes espaciales de la componente de luminancia Y, dos se basan en características extraídas del vector formado por las dos componentes de crominancia (Cb, Cr) y uno se basa en las características del contraste y la información temporal absoluta, extraídos de la componente de luminancia Y. La selección de los parámetros de calidad de vídeo se determinó por los criterios de maximizar la correlación entre la calidad perceptual estimada (obtenida del modelo) y la medida subjetivamente. El modelo general produce valores de salida (VQM_G) que van de cero (no hay degradación percibida) a uno (máxima degradación percibida) aproximadamente. Para llevar los resultados a la escala de calidad continua de doble estímulo (DSCQS), se debe multiplicar por 100.

Dentro de este modelo, una "característica de calidad" se define como la cantidad de información asociada a una subregión espacio-temporal de una secuencia de vídeo (original o procesada). Las secuencias de características producidas son función del espacio y del tiempo. Comparando las características extraídas del vídeo procesado (degradado) con las extraídas del vídeo original, puede calcularse un conjunto de parámetros de calidad, que son indicativos de las variaciones perceptuales de la calidad de vídeo.

En forma genérica, los cálculos de las características se realizan de acuerdo con los siguientes pasos:

1. Aplicar un filtro perceptual (Opcional)
2. Dividir la secuencia de vídeo en regiones S-T (espacio – temporales)
3. Extraer las características de cada región S-T (espacio – temporales)
4. Aplicar un umbral de perceptibilidad (Opcional)

Las regiones S-T corresponden a un área rectangular de la imagen, tomada en cierta cantidad de campos consecutivos, como se esquematiza en la Figura 5.8. Por ejemplo, pueden tomarse áreas de 8 x 8 píxeles, y una duración de 0.20 segundos (6 cuadros de video en NTSC o 5 cuadros en PAL) [41]

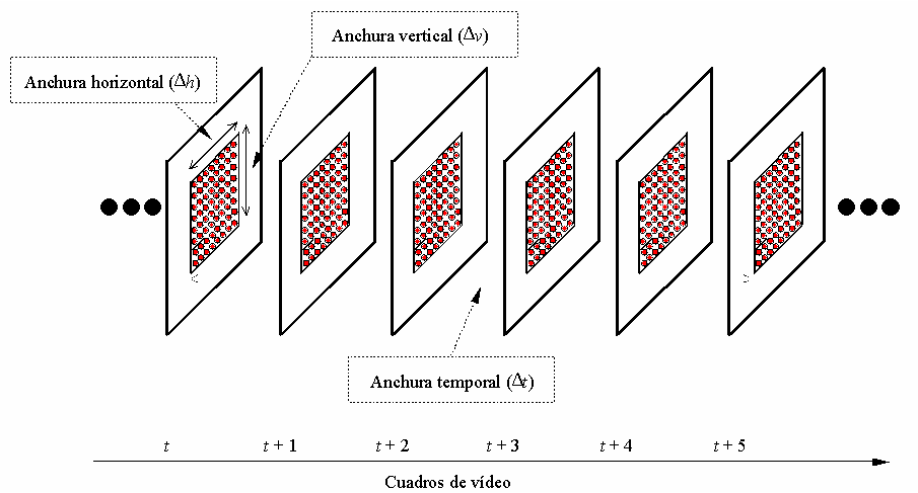


Figura 5.8

A continuación se describen conceptualmente las características del video utilizadas por este modelo para realizar sus predicciones acerca de la calidad percibida.

Características basadas en gradientes espaciales
Se pueden utilizar características derivadas de gradientes espaciales para caracterizar las distorsiones perceptuales de los bordes. Las componentes de luminancia *Y* de las secuencias de vídeo original y procesada (degradada) se filtran por medio de filtros de realce de bordes horizontales y verticales. A continuación, estas secuencias de vídeo filtradas se dividen en regiones S-T de las que se extraen las características resumidas, que cuantifican la actividad espacial en función de la orientación angular. Luego, un proceso aplica umbrales de perceptibilidad.

Características basadas en la información de crominancia
Se calcula un vector que representa los valores medios de las componentes Cb y Cr dentro de la región S-T, ponderando en un factor de 1.5 la componente Cr, según la siguiente ecuación:

$$f_{coher\_color} = \left(\overline{C}_b, 1.5 \cdot \overline{C}_r\right) \qquad (5.11)$$

Un tamaño de región S-T de 8 píxeles horizontales × 8 líneas verticales × (1 a 3) cuadros de vídeo produce un vector de característica de crominancia aceptable para estos cálculos (en realidad, quedan 4 x 4 píxeles de crominancia, debido al submuestreo de las señales de color)

Características basadas en la información de contraste
Las características que miden información de contraste localizada son sensibles a degradaciones de calidad tales como la borrosidad (por ejemplo, pérdida de contraste) o una eventual ganancia de contraste. Una característica de contraste localizada se calcula en este modelo para cada región S-T como la desviación estándar de la luminancia Y

Características basadas en ATI (Absolute Temporal Information)
La "ATI" cuantifica la cantidad de movimiento en una escena de video. Las características que miden distorsiones en el flujo de movimiento son sensibles a degradaciones de calidad tales como la omisión o repetición de cuadros (pérdida de movimiento). Se calcula una característica de ATI para cada región S-T generando primero una secuencia de vídeo en movimiento que sea el valor

absoluto de la diferencia entre los cuadros de vídeo consecutivos en los instantes t y $t-1$, y calculando a continuación la desviación típica en la región S-T.

Características basadas en el producto vectorial del contraste y la ATI
Los efectos de enmascaramiento hacen que la perceptibilidad de las degradaciones espaciales puede verse influida por la cantidad de movimiento presente. Análogamente, la perceptibilidad de las degradaciones temporales puede verse influida por la cantidad de detalle espacial presente. Puede utilizarse una característica derivada del producto vectorial de la información de contraste y de la información temporal absoluta (ATI) para explicar, al menos en parte, estas interacciones.

Los parámetros de calidad anteriormente descritos se calculan en primer lugar para cada región S-T de la señal original, y de la señal procesada (degradada). Los parámetros de las regiones S-T forman matrices tridimensionales que comprenden un eje temporal y dos dimensiones espaciales (horizontal y vertical). A continuación, para cada cuadro (correspondiente a un instante t del eje temporal) se agrupan las degradaciones utilizando una función de agrupamiento espacial. El agrupamiento espacial produce una historia temporal de los valores de los parámetros. Esta historia temporal de los valores de los parámetros se agrupa luego temporalmente por medio de una función de agrupamiento temporal, dejando un único parámetro de cada característica para toda la secuencia de video (típicamente, del orden de 10 segundos). Cada parámetro *p* agrupado temporalmente puede ajustarse a escala en razón de las relaciones no lineales entre el valor del parámetro y la calidad percibida. Finalmente, se realiza una suma ponderada de los parámetros para tener la estimación de la calidad perceptual.

Es de hacer notar que las investigaciones realizadas por los desarrolladores del modelo, indican que las funciones de agrupamiento espacial óptimas deben conllevar algún tipo de procesamiento que tenga especial énfasis en los casos más desfavorables. Esto se debe a que las degradaciones localizadas tienden a atraer la atención del observador, haciendo que la parte más desfavorable de la imagen sea el factor predominante en la decisión de calidad subjetiva.

5.2.2. RRNR-TV (Reduced Reference/No reference TV)

Al igual que el FR-TV, el objetivo principal de las pruebas planificadas por el VQEG es evaluar las métricas de calidad de video (VQMs) propuestas en diversos modelos, con los resultados subjetivos obtenidos mediante SSCQE. Se prevé tomar muestras de MOS cada 0.5 segundos en esta etapa de pruebas subjetivas El último documento disponible acerca del plan de pruebas de RRNR-TV es del 28 de setiembre de 2006 [45].
Las secuencias de prueba incluirán formatos de 50 Hz y 60 Hz, y se admitirán para los modelos RR canales de referencia de 10 kb/s, 56 kb/s y 256 kb/s. Los modelos NR no requieren canales de referencia, y basan sus medidas únicamente en la señal de video degradada.
Asimismo, se prevén codificaciones MPEG-2, H.264 o VC1, y velocidades de 1 Mb/s a 6 Mb/s

5.2.3. MM (MultiMedia)

En este contexto, “Multimedia” se define como una aplicación que puede combinar texto, gráficos, video y sonido dentro de un mismo paquete. Estas aplicaciones incluyen, por ejemplo, sistemas de video conferencias. Las herramientas a evaluar por el grupo MM pueden ser utilizadas tanto en ambientes de laboratorio (con modelos FR) como en ambientes operacionales, con modelos RR y NR. El análisis será basado en medidas del tipo DMOS para los modelos FR y RR, y medidas de MOS para el modelo NR.

Se espera que las pruebas subjetivas se realicen en forma separada para los diferentes tipos de “monitores” utilizados. Por ejemplo, se prevén pruebas específicas para dispositivos móviles como PDAs, independientes de las pruebas realizadas en ordenadores de escritorio, ya que las

características visuales de estos dispositivos son muy diferentes. Por lo tanto, es posible que del resultado de las pruebas, surjan modelos más adecuados a cada tipo de dispositivo.

En el caso de RR, se admitirán las siguientes velocidades para el canal de referencia, dependiendo del dispositivo utilizado:

- PDA/Mobile (QCIF): (1 kb/s, 10 kb/s)
- PC1 (CIF): (10 kb/s, 64 kb/s)
- PC2 (VGA): (10 kb/s, 64 kb/s, 128 kb/s)

La última versión disponible del plan de pruebas del grupo MM es la 1.16, del 7 de febrero de 2007 [46]

5.2.4. HD-TV (High Definition TV)

En este caso, el objetivo es evaluar modelos de calidad perceptual enfocados en aplicaciones de televisión digital de alta definición (HD-TV). Esto incluye broadcasting, video a demanda y TV por cable o satélite, presentados en pantallas de ordenadores, o en monitores de televisión de alta definición.

Se espera que las pruebas subjetivas se realicen en forma separada para los distintos formatos de HDTV.

Se prevén velocidades de 1 Mb/s a 30 Mb/s, con codificación AVC/H.264/MPEG4 part 10. También se prevé incluir errores de transmisión en las pruebas, incluyendo pérdida de paquetes, demoras, jitter, errores de bits y errores “en el aire” para comunicaciones inalámbricas.

Al igual que en el resto de las pruebas, se admiten modelos FR, RR y NR. No está establecido al momento, los anchos de banda preemitidos para los canales de referencia de los modelos RR.

5.3. Otras propuestas de modelos FR

La mayoría de los métodos presentados utilizan una aproximación como la que se ilustra en la Figura 5.9, donde las señales original y degradada son preprocesadas (por ejemplo, alineadas espacial y temporalmente, se aplica alguna transformada de luminancia o crominancia, etc.) y luego descompuesta en diferentes “canales”, sobre los que se calcula un “error” según algún tipo de métrica. Estos errores, generados para cada “canal” son sometidos a reglas de enmascaramiento y luego ponderados e integrados (sumados) para obtener finalmente una única medida de calidad.

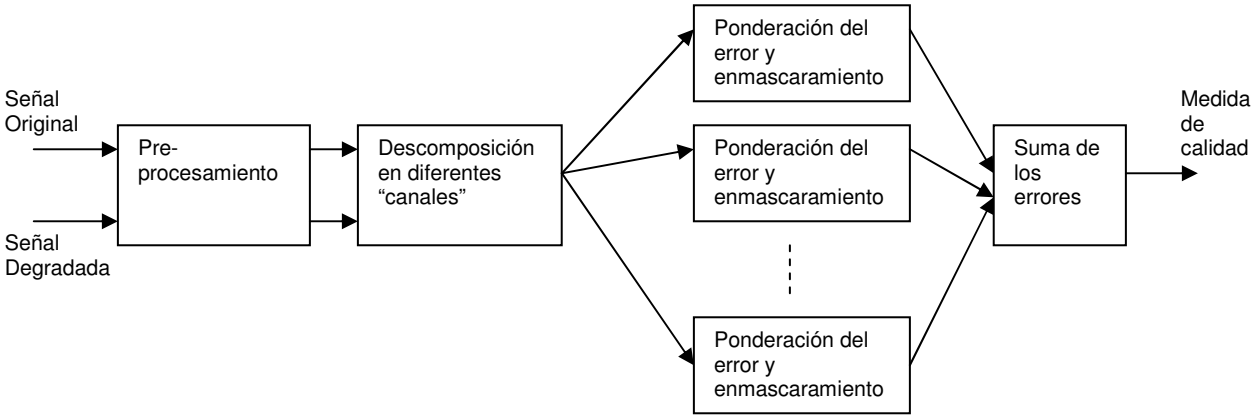


Figura 5.9

En [47] se presenta una serie de críticas genéricas a este tipo de modelos, y se propone, como contraparte, una "nueva filosofía" en el diseño de métricas de medida de la calidad perceptual de video. Esta nueva filosofía se basa en que *"la función principal de los ojos humanos es extraer información estructural del campo visual, y el sistema visual humano está altamente adaptado para este propósito. Por lo tanto, una medida de las distorsiones estructurales debería ser una buena aproximación a la distorsión percibida"* [47]. Los autores de esta propuesta muestran como grandes errores absolutos en algún canal, entre la señal original y la degradada, no siempre implican grandes distorsiones estructurales. Se propone en el mencionado artículo, como medida de calidad perceptual, el índice Q, definido como

$$Q = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \cdot \frac{2\overline{x}\overline{y}}{(\overline{x})^2 + (\overline{y})^2} \cdot \frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \qquad (5.12)$$

donde x representa a la señal original, y a la señal degradada y σ a la varianza o covarianza.

El primer término de la ecuación corresponde con la correlación entre la señal original y la degradada. El segundo término mide qué tan cercanos entre sí son los valores medios. El tercer término indica qué tan similares son las varianzas de ambas señales.

El método aplica el factor Q a ventanas móviles de 8 x 8 píxeles, y luego promedia todos los valores para obtener el índice general de cada imagen. El método es sencillo, y la métrica Q tiene una interpretación clara. Aplicando este método a la Figura 5.1, se obtienen los valores de Q=0.3709 para "Tiffany", Q=0.4606 para "Lago" y Q=0.7959 para "Mandril", lo que a priori se corresponde con la calidad percibida mucho mejor que la métrica MSE, casi idéntica en las tres imágenes.

SSIM (Structural SIMilarity)

Basado en el concepto anterior, el índice denominado SSIM (Structural SIMilarity) [48] fue propuesto recientemente. En este modelo, la similitud entre dos imágenes depende de la similitud entre sus luminancias, contrastes y estructuras. La similitud general (S) entre una imagen x y una imagen y es, según este modelo, una función de las similitudes de estas tres componentes:

$$S(x, y) = f\big(l(x, y), c(x, y), s(x, y)\big) \qquad (5.13)$$

donde l, c y s representan las funciones de similitud de luminancia, contraste y estructura respectivamente.

El modelo presenta la definición de cada una de estas funciones, teniendo en cuenta los siguientes criterios:

- La función de comparación de luminancia l debe tener en cuenta la ley de Weber
- La función de similitud general S debe:
  - ser simétrica ($S(x,y) = S(y,x)$)
  - ser acotada a $S(x,y) <= 1$
  - tomar su máximo valor 1, si y solo si x = y

Se sugiere en este modelo una función S donde se utiliza el producto de cada función de comparación, ponderado por un exponente, de la forma

$$S = l^{\alpha} \cdot c^{\beta} \cdot s^{\gamma} \qquad (5.14)$$

En particular, tomando el valor 1 para todos los exponentes, el índice final SSIM propuesto toma la forma

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \qquad (5.15)$$

Siendo μ valores medios, σ varianza o covarianza, y C_1 y C_2 dos variables de ajuste.

Este índice puede ser calculado en ventanas móviles de, por ejemplo, 8 x 8 píxeles, obteniendo un mapa de valores SSIM, los que luego pueden ser promediados para obtener el índice de la imagen completa (MSSIM). Si se conocen las áreas de la imagen de fijación de la vista, se pueden ponderar los índices en esta zona, de manera que pesen más en la suma total.

En [24] se presentan resultados obtenidos con este modelo, dónde se puede ver que el índice MSSIM es alrededor de un 8% mejor que el PSNR, utilizando el coeficiente de correlación de Pearson, para imágenes JPEG. Sin embargo, esta mejora no parece realmente sustantiva, y está por debajo de las mejoras logradas en los modelos evaluados por el VQEG, presentados en el capítulo 5.2.1.

5.4. Modelos NR

En los modelos FR se mide básicamente similitudes o diferencias entre una imagen o video de referencia, y otra imagen o video degradadas. Los modelos NR, en contraste, intentan predecir la calidad subjetiva de una imagen o video basándose únicamente en el análisis de la señal recibida. Esta tarea, realizada en forma sencilla y natural por las personas, es sumamente difícil de realizar de manera automática. No existe al momento un modelo perceptual estandarizado por organismos internacionales que aplique al caso de NR (auque, como se vio, el VQEG está trabajando en el tema). Las aproximaciones existentes consisten en realizar la estimación de calidad en base a las medidas de algún conjunto de “degradaciones tipo”, que en la mayoría de las situaciones afectan a la calidad percibida.

El modelo NROQM (No Reference Objective Quality Metric) [49] se basa en estimar el contenido de 6 características del video (contraste, ruido, clipping, ringing, efecto de bloques y definición), y obtener un valor de calidad perceptual basado en las medidas de cada uno de éstas características. El índice de calidad obtenido es una combinación lineal de los índices de cada

efecto, los que se calculan en base a fórmulas heurísticas, basadas en aproximar las medidas objetivas con pruebas subjetivas de calibración.
En los resultados presentados en [49] se indica que la correlación obtenida entre el NROQM y las pruebas subjetivas fue de 0.85, mientras que la correlación con el PSNR de 0.399. Estos resultados no condicen con el informe del VQEG, donde la correlación media obtenida con la métrica PSNR está cercana al 80% promediando un gran número de casos con diferentes tipos de videos. Igualmente cabe resaltar que en el caso del NROQM no se dispone de la señal de referencia para la estimación, por lo que el valore de 0.85 presentado parece ser sumamente aceptable.

Existen algunos productos comerciales que utilizan el modelo NR para la predicción de la calidad perceptual de Video. El modelo PQoS de Genista Corporations, se basa en la estimación del efecto de bloques, la borrosidad, y el "jerkiness" (interrupciones) para estimar la calidad percibida. En [50] se realiza una comparación de este método contra pruebas subjetivas, mostrando una correlación media de un 78% entre el MOS subjetivo y el estimado por PQoS.

Dado que la mayoría de los modelos NR toma el efecto de bloques como uno de los principales componentes que afectan a la calidad perceptual, es importante disponer de índices numéricos que lo estimen. Varios métodos se han propuesto, como se describe a continuación.

En [51] se propone estimar el efecto de bloques basado simplemente en medir las diferencias horizontales y verticales entre las filas y columnas en los bordes de los bloques.

En [52] se propone un algoritmo basado en la correlación cruzada de imágenes sub-muestreadas. Se generan cuatro imágenes, en las que cada una contiene los píxeles de una de las cuatro esquinas de cada bloque de 8 x 8, y se realizan correlaciones entre éstas imágenes y otras imágenes derivadas de la selección de píxeles específicamente seleccionados de cada bloque.

Otra idea, propuesta en [53], modela las imágenes como la suma de dos componentes: una imagen sin distorsiones, y una imagen puramente de bloques. El algoritmo consiste en aplicar a la señal degradada una transformada FFT de una dimensión en forma horizontal y vertical, de manera de identificar los picos producidos por el efecto de bloques, los que se encuentran, si existen, en lugares bien definidos del espectro. A su vez, el espectro de la señal original (sin distorsión de bloques) es estimado mediante la aplicación de filtros adecuados. La diferencia de ambas señales, corresponde al efecto de bloques, y puede evaluarse numéricamente.

Estos tres métodos han sido comparados en [54] .

Es de hacer notar que los modelos NR, detallados anteriormente, se basan en la existencia de algunos tipos de degradaciones específicas, producidas por las técnicas de codificación. El efecto de bloques es resultado de la aplicación de la transformada DCT a pequeños bloques de imagen. La transformada DWT, por el contrario, no produce efectos de bloques. Si bien por ahora esta transformada se utiliza típicamente para imágenes fijas (ver JPEG2000 en 3.1), es posible que en un futuro pueda utilizarse también para video, lo que llevaría a que los modelos NR basados en la estimación del efecto de bloques u otras degradaciones específicas del tipo de codificación, no sean adecuadas.

6. Efecto de los problemas de redes IP en la calidad de Video

Como se ha visto en los capítulos anteriores, el mayor esfuerzo en la medida objetiva de la calidad perceptual de video se ha realizado hasta el momento en el ámbito de aplicaciones de televisión, y en modelos FR, es decir, partiendo de la base de que se dispone de la señal original de referencia.

La transmisión de video sobre redes de paquetes, y en particular, sobre la Internet, presenta características esencialmente diferentes a la de la difusión de TV por las vías clásicas. Se utilizan rangos de anchos de banda variables, hay congestión y pérdida de paquetes, y típicamente la observación se realiza desde distancias más cortas, y generalmente en pantallas más pequeñas. Si bien el VQEG está en proceso de estudio de modelos para estos ambientes (por ejemplo, en el proyecto Multimedia), estos modelos recién están en su etapa temprana de estudio.

En esta sección se describirán los factores específicos de las redes IP que afectan a la calidad de video, y se presentarán referencias a estudios y propuestas recientes sobre el tema.

6.1. Efecto de la pérdida de paquetes en la calidad de video

La pérdida de paquetes en las redes IP afecta a la calidad percibida de video. En la Figura 6.1, tomada de [55], se muestra como la pérdida de un paquete puede propagarse, afectando no solo a la información de video contenida en dicho paquete, sino a otras partes del mismo o diferentes cuadros. Típicamente, y dado que la codificación se realiza en forma diferencial, la pérdida de un paquete afectará a todos los bloques siguientes en la misma fila ("slice"). Si el paquete perdido corresponde además a un cuadro de referencia (I), también se verán afectados los cuadros predictivos, posteriores o anteriores, propagándose el error en el tiempo.

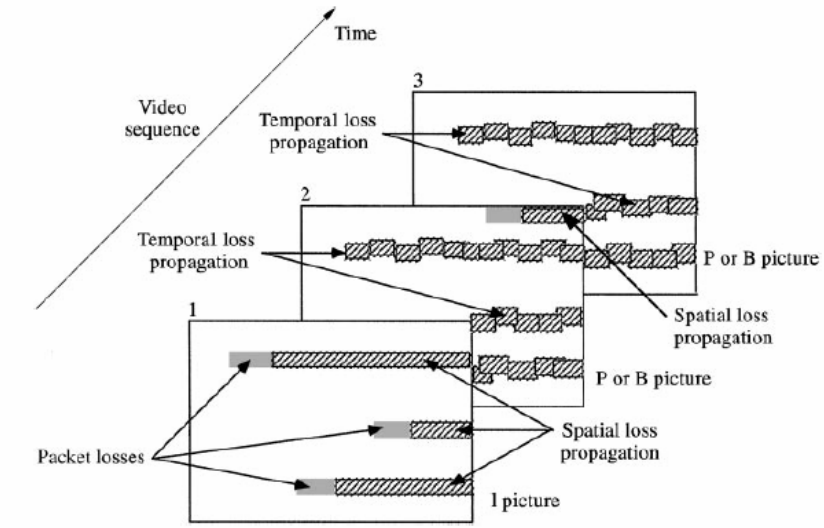


Figura 6.1

Existen técnicas de cancelación de paquetes perdidos, las que tratan de reconstruir la información perdida en base a información disponible. Por ejemplo, reemplazando los píxeles perdidos por los mismos valores de cuadros anteriores.

Varios trabajos se han realizado, estudiando la manera en que la calidad de video se ve afectada por la pérdida de paquetes. En [56] se propone estimar el MSE del video, en base a 3 técnicas diferentes, para los casos de pérdida de paquetes: En la primer técnica ("Full Parse"), se inspeccionan todos los paquetes y se obtiene información relevante del video, sin llegar a decodificarlo. En la segunda técnica ("Quick Parse"), se inspeccionan los paquetes con menor detalle, llegando hasta los cuadros y líneas, pero sin llegar a la profundidad de la técnica anterior. Este proceso es naturalmente más rápido y requiere menos procesamiento. Finalmente, una tercer técnica ("No Parse") se basa únicamente en la medida de pérdida de paquetes, sin realizar otras inspecciones ni decodificar el video. Para cada una de estas técnicas se proponen formas de estimar el MSE. Los autores han realizado un prototipo para MPEG-2, y concluyen que los

estimadores se aproximan a los valores reales del MSE, siendo el Full Parse el método más aproximado, y el No Parse el más alejado. El modelo propuesto es del tipo NR, ya que no requiere de la referencia de video original para su estimación.

En [57] se muestra que no alcanza con conocer el porcentaje de pérdida de paquetes para estimar como se ve afectada la calidad de video percibida. Dependiendo de diversos factores, la pérdida de un paquete determinado de video puede o no afectar la calidad percibida. Por ejemplo, en imágenes casi estáticas, el video perdido puede ser reconstruido en base a imágenes anteriores, casi sin pérdida de calidad, lo que lleva a que la pérdida de un paquete sea prácticamente imperceptible. Algo similar sucede cuando la pérdida solo afecta a un cuadro. De esta manera, se propone un "clasificador de paquetes perdidos", que, en base a un algoritmo, decide si el paquete perdido afectará o no a la calidad percibida del video. El algoritmo toma en cuenta cuantos cuadros se verán afectados por la pérdida del paquete, la movilidad de la imagen y su varianza y el error introducido medido con MSE, entre otros factores. Se proponen cuatro algoritmos diferentes, los que varían en la complejidad de cálculo. Según los autores, con el algoritmo más complejo se logra un 93% de aciertos en la decisión de si el paquete perdido afectará o no a la calidad percibida. En base a esto, se propone el uso de un coeficiente que toma en cuenta únicamente la tasa de paquetes perdidos que afectan a la calidad de video, al que definen como VPLR (Visible Packet Loss Rate), en lugar del clásico PLR (Packet Loss Rate).

La idea de detectar como afecta la posible pérdida de cada paquete en la calidad perceptual es explorada en [58], donde se propone un método que garantice una calidad perceptual constante en la recepción de un flujo de video. La idea en este caso es marcar solo ciertos paquetes específicos con mayor prioridad (asumiendo una red con soporte para Diff Serv), de manera de "garantizar" su llegada a destino, sin inundar a la red con todos los paquetes de video marcados como prioritarios, sino solo con los paquetes cuya pérdida afecten especialmente la calidad percibida. El algoritmo se basa únicamente en la calidad deseada (PSNR) y la tasa de pérdida de paquetes (PLR).

Una idea similar es presentada en [59], donde se presenta una métrica para priorizar ciertos paquetes de video, en base a la estimación de la distorsión percibida en caso de la pérdida o llegada fuera de tiempo de cada paquete.

Dado que las perdidas de paquetes se dan generalmente en ráfagas, la calidad puede verse fuertemente degradada por la pérdida de varios paquetes consecutivos, correspondientes a cuadros consecutivos. En [60] se propone una técnica que consiste en reagrupar los cuadros que se envían, generando un buffer en el codificador de 3 GOPs, y reagrupando el orden en el que se envía la información. Con esto se logra difundir los paquetes perdidos entre varios cuadros separados en el tiempo, y de esta manera, según los autores, mejorar la calidad percibida. En el método propuesto se evalúa el caso de MPEG-4 con un GOP de 12 cuadros del tipo IBBPBBPBBPBBI. Sin embargo, no se dispone en el artículo de medidas de la calidad percibida objetiva ni subjetiva, por lo que no se puede tener una medida de la mejora del método que proponen.

Si bien varios trabajos se han realizado acerca de la degradación del video debida a la pérdida de paquetes, el tema está aún abierto, y no hay aún estándares ni trabajos sistemáticos de comparación de diferentes modelos.

6.2. Efecto de la demora / Jitter

El receptor debe recibir los paquetes a decodificar a intervalos constantes, para poder regenerar de forma adecuada la señal original. Dado que el *jitter* es inevitable en las redes de paquetes, los receptores disponen de un *buffer* de entrada, con el objetivo de "suavizar" el efecto de la variación de las demoras. Este buffer recibe los paquetes a intervalos variables, y los entrega a intervalos constantes.

Es de hacer notar que este buffer agrega una demora adicional al sistema, ya que debe retener paquetes para poder entregarlos a intervalos constantes. Cuánto más variación de demoras (jitter) exista, más grande deberá ser el buffer, y por lo tanto, mayor demora será introducida al sistema. Las demoras son indeseables, y tienen impacto directo en la experiencia del usuario, sobre todo en contenidos de tiempo real (por ejemplo, distribución de eventos deportivos en línea) y en aplicaciones conversacionales (por ejemplo video telefonía, o video conferencias).
Se hace necesario disponer de mecanismos que minimicen el tamaño de los jitter-buffers, pero que a su vez, no comprometan la calidad debida a la perdida de paquetes que no han llegado a tiempo.

En [61] se propone un método dinámico al que llaman AMP (Adaptive Media Playout), en el que se cambia la velocidad de reproducción del medio (video / audio) dependiendo de la condición del canal de transmisión, y logrando de esta manera reducir el tamaño del jitter-buffer. Los autores del artículo referenciado indican que aumentar o disminuir la velocidad de ciertas partes del contenido hasta en un 25% es subjetivamente mejor que aumentar las demoras totales o tener interrupciones. En la Figura 6.2, tomada de [61] se muestra un ejemplo en el que se ve como, aplicando la técnica AMP propuesta, se puede bajar a la mitad el tiempo del comienzo de la reproducción. En este ejemplo, se dispone de un buffer de 10 segundos. En condiciones normales, éste es el tiempo que se debería esperar para comenzar a reproducir el medio en el receptor. Aplicando AMP, la reproducción comienza a los 5 segundos, pero a una velocidad de reproducción ligeramente menor a la que debería durante los primeros 25 segundos de reproducción.

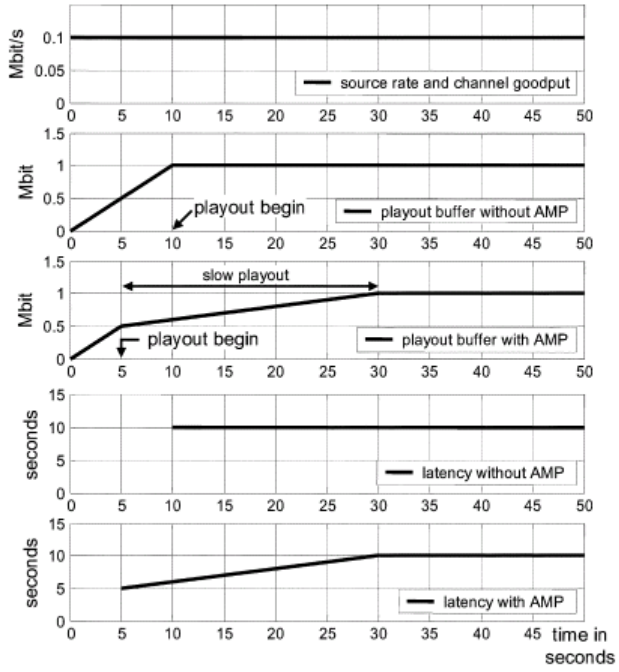


Figura 6.2

En el proyecto Multimedia del VQEG se proponen realizar pruebas con demoras entre 2 milisegundos y 5 segundos, y pérdidas de paquetes impulsivas en el rango de 0 a 50%.

7. Futuras líneas de investigación

Muchos aspectos referentes a la calidad perceptual de video están aún en proceso de comprensión, y requieren de mayores investigaciones. Es por eso que en esta sección se detallan

resumidamente algunos de los aspectos que podrían dar lugar a nuevas líneas de investigación, las que podrían ser parte de un trabajo de Tesis Doctoral.

7.1. Modelo de visión humano

No se dispone al momento de un modelo completo del sistema de visión humano. Si bien muchos aspectos son conocidos, en última instancia no se conoce como es el mecanismo por el cual el cerebro procesa la información y como la utiliza para sus procesos cognitivos. Por lo tanto, los algoritmos de predicción de calidad perceptual basados en el sistema de visión humano están aún en sus etapas preliminares y son tema de investigación abierta.
Comprender estos aspectos sería muy importante para generar un sistema automático que tenga en cuenta los aspectos preceptuales.

Los trabajos realizados al momento intentan generar un modelo que prediga, con la mayor exactitud posible, la calidad "percibida subjetivamente". La medida de la calidad subjetiva se realiza, de acuerdo a las recomendaciones actuales ([19] [20]), calificando con un *único* índice el material a prueba. Cabe preguntarse si esto es suficiente, es decir, si basta un único número para calificar la calidad percibida de, por ejemplo, un contenido multimedia. Es posible que futuras investigaciones demuestren se pueden obtener mejores modelos con el uso de varios índices, en un sistema multidimensional.

7.2. Estimación de la calidad perceptual de video en base a parámetros de la red

Un área de particular interés es la transmisión de video y multimedia sobre redes IP, y la evaluación de la calidad perceptual objetiva en base a medidas de los parámetros de desempeño de la red. En audio se dispone del "Modelo E", estandarizado en la recomendación ITU-T G.107 [62], el que permite predecir en forma objetiva, y dentro de ciertos márgenes de error, la calidad percibida en base a los codecs utilizados y las características de la red IP (por ejemplo, demoras, jitter, pérdida de paquetes). Si bien este modelo ha sido criticado por su inexactitud bajo ciertas circunstancias, no cabe duda que es una aproximación muy importante y sobre todo, que brinda un sencillo marco de referencia. No existe, hasta el momento, un estándar similar para video, y hay aquí un camino de investigación.

En éste análisis, la calidad de video deberá ser primeramente categorizada en función del codec utilizado, la resolución y el ancho de banda. Como se vio en la Figura 3.4 de la sección 3.3.3, hay una curva de calidad vs ancho de banda, dado un codec y una secuencia de video. Sin embargo, en la literatura actual tanto comercial como académica, estas gráficas presentan típicamente la calidad medida con la métrica PSNR, que, como ya se vio, no se corresponde con la calidad perceptual. Un primer aporte debería llevar a una evaluación de la calidad vs el ancho de banda en base a alguno de los modelos de FR ya estandarizados. Adicionalmente, este análisis se puede realizar para un conjunto importante de secuencias de video, y seguramente se pueda llegar a una "gráfica promedio", la que dependerá del tipo de señal (con mucho o poco movimiento, con mucha o poca textura, etc.) Este camino conduce a hallar el equivalente al parámetro I_e del Modelo E, es decir, a categorizar los codecs en base a como afectan la calidad percibida.

El aporte a la degradación de los problemas típicos de las redes de datos (pérdida de paquetes, jitter, demoras) podrá ser evaluado luego, en forma ordenada y controlada, y siempre utilizando como métrica de calidad perceptual alguno de los modeles FR ya estandarizados.

En el ámbito corporativo, las aplicaciones de video telefonía y video conferencias están comenzando a desarrollarse. Por ejemplo, está teniendo fuerte impulso comercial el producto "Microsoft Live Communication Server", y en pocos meses estará disponible el "Microsoft Office

Communications Server". Ambos disponen de aplicaciones de video telefonía y video conferencias en el escritorio. Asimismo, estas aplicaciones se vinculan con las "clásicas" PBX, las que pueden cursar llamadas de voz y video. Hace algunos años, cuando comenzaron a aparecer las aplicaciones de VoIP (Voz sobre IP), conllevaron consigo la caracterización de las redes ("network assessment) corporativas, para evaluar su aptitud para soportar la Voz sobre IP. El Modelo E, junto con las directivas de la TIA/EIA [63] fueron, y siguen siendo, la base de estimación a nivel corporativo para calificar el estado de la red al momento de incorporar aplicaciones de VoIP. Con el advenimiento de las comunicaciones de video, una nueva caracterización de las redes corporativas será necesaria, ya que los requerimientos de anchos de banda, demoras, porcentaje de pérdidas de paquete, etc. deberán adecuarse a las necesidades de ésta nueva aplicación. Disponer de un modelo simple, como el Modelo E, será fundamental para esta tarea. Es por ello que se entiende conveniente investigar la posibilidad de avanzar en el camino de crear este modelo, dentro de las limitaciones que se vean necesarias, pero como una aproximación de medida de calidad percibida basada en parámetros fácilmente mensurables, como el codec, el ancho de banda y los parámetros de la red.

7.3. Estudio detallado y aplicabilidad de técnicas FR

Los mayores avances hasta el momento son en el ámbito de modelos del tipo FR (Full Reference), donde el VQEG ha realizado un extenso trabajo, evaluando y comparando diversos modelos propuestos. Sin embargo, es interesante notar que aún los modelos que presentaron mejor desempeño, y que fueron elegidos como estándares de ITU, presentan solo un desempeño "levemente" mejor que el sencillo PSNR (del orden de 20% mejor, como se vio en 5.2.1), y ninguno de ellos llega al "límite teórico", dado por la varianza de las pruebas subjetivas. A su vez, cada uno de estos modelos se basa en características preceptuales muy diferentes. Por ejemplo, para mencionar solo dos casos, el modelo de Yonsei University proporciona un método de medición objetiva de la calidad de video basado exclusivamente en la degradación en el entorno de los bordes, mientras que el modelo de la NTIA toma parámetros de una amplia gama de degradaciones tales como la borrosidad, el efecto de bloques, el movimiento entrecortado / innatural, el ruido y los bloques con errores. Es interesante notar que ambos pueden predecir de forma estadísticamente equivalente la calidad perceptual promedio subjetiva, aún modelando aspectos cognitivos muy diferentes. De alguna manera esto denota el poco conocimiento real que se tiene del sistema visual humano.

Los resultados del estudio del VQEG a modelos FR muestran diferencias según la cantidad de líneas de la imagen (525 o 625). No se evidencia, a priori, ninguna razón por la que la calidad percibida deba ser diferente según la cantidad de líneas de la imagen (por ejemplo, no hay diferencias preceptuales apreciables entre los sistemas PAL y NTSC). Algunos modelos han presentado mejores resultados para 525 líneas, mientras que otros lo han hecho para 625 líneas. Esto podría ser indicio de que hay otros factores que estén afectando al resultado de estas pruebas, y podría ser interesante analizarlo más detenidamente.

Evidentemente, la capacidad de proceso requerida para implementar estos algoritmos es notoriamente diferente entre sí, y este aspecto, no evaluado por el VQEG será de real importancia al momento de las implementaciones prácticas. Una interesante tarea de investigación será la comparación detallada de la capacidad de proceso requerida por cada modelo, buscando algoritmos que optimicen sus implementaciones, y una evaluación de si son posibles de implementar en sistemas de tiempo real.

7.4. Aplicación de técnicas FR a modelos NR

Los modelos del tipo RR y NR están aún más inmaduros que los del tipo FR. Si bien ha habido propuestas de algoritmos basados en estos modelos, aún no se han realizado las comparaciones

sistemáticas (aunque ya están propuestas por el VQEG). A su vez, el paradigma de estos modelos (en especial del NR) es esencialmente diferente al de los modelos FR. Varios modelos FR intentan explotar las características genéricas del sistema visual humano, lo que de alguna manera los independiza de los tipos de degradaciones presentes en el video (y por lo tanto, del tipo de codificación y digitalización). Por otra parte, los sistemas NR explotan menos los aspectos del sistema visual humano, y se basan principalmente en cuantificar diferentes tipos de degradaciones típicas del video codificado digitalmente. Es decir, los modelos NR están mucho más ligados al tipo de codificación, aplicación y transmisión utilizados, ya que, por lo menos en gran parte de los modelos propuestos hasta el momento, se parte de la base de que existirá cierto tipo específico de degradaciones (por ejemplo efecto de bloques, ringing, etc.). Una interesante línea de investigación consistirá en evaluar la posibilidad del uso de las técnicas ya probadas en modelos FR aplicadas a modelos NR. Si bien esto no parece posible a priori, por no disponer de la señal de referencia, el autor del presente trabajo entiende que este camino es viable. Un método propuesto para el caso de audio [64], sugiere poner un punto de prueba en el receptor, y sustituir en cada paquete el "contenido" del medio por un contenido de referencia conocido. Es decir, la idea subyacente consiste en predecir la calidad percibida de una señal cualquiera, en base a una medida FR sobre una señal *conocida* que se vea afectada por exactamente las mismas degradaciones que la señal a evaluar. Esta señal *conocida* deberá ser "equivalente" a la señal que se desea medir en lo que respecta a la calidad perceptual.

Sustituyendo el contenido de los paquetes recibidos por el video "equivalente", se puede tener una señal de referencia conocida, y la misma señal sufriendo "las mismas" degradaciones que la señal a evaluar. Esta idea requiere del estudio de varios aspectos para evaluar su viabilidad para video. Es conocido, como se vio en el capítulo 6, que la visibilidad de cada paquete perdido depende del tipo de codificación, de la posición del contenido del paquete en la pantalla o incluso del tipo de contenido del video. Probablemente los efectos de enmascaramiento en imágenes con mucho movimiento, o con mucha textura (actividad espacial) hagan menos visibles las degradaciones frente a imágenes estáticas, o con poca actividad espacial. Para aplicar este método será seguramente necesario caracterizar el tipo de contenido recibido, y elegir de esta manera el video "equivalente" más apropiado, con la hipótesis de que en tiempos medios (por ejemplo, algunos minutos), la evaluación de la calidad percibida subjetiva serán las mismas entre el video real y el video "equivalente".

La caracterización del contenido es en si mismo otro interesante camino de investigación que puede resolverse con algoritmos de "caracterización de tipo de contenido", o con el envío de este tipo de información desde el emisor, convirtiendo quizás el modelo en uno del tipo RR, pero haciendo necesaria la modificación de los codecs, ya que deberían enviar esta información dentro del contenido.

Si este camino conduce a buenos resultados, se podrán reutilizar los mismos modelos FR ya estandarizados por la ITU, y usarlos en modelos NR, lo que supondría un avance sustantivo en la materia.

8. Conclusiones

En este trabajo se ha presentado un detalle del "estado del arte" en el tema de medida objetiva de calidad perceptual de video. El área es de gran interés y actualidad, y existe un número importe de investigadores que están trabajando en el tema. Sin embargo, aún hay un gran camino por recorrer, y varios aspectos están aún abiertos.

Los mayores estudios al momento, y los únicos recientemente estandarizados por organismos internacionales ([41]), refieren al estudio de degradaciones para aplicaciones de TV, con modelos FR (Full Reference). El VQEG ha realizado un extenso trabajo en éste área, evaluando diversos modelos propuestos. Sin embargo, es interesante notar que aún los modelos que presentaron

mejor desempeño, y que fueron elegidos como estándares de ITU, presentan solo un desempeño "levemente" mejor que el sencillo PSNR (del orden de 20% mejor, como se vio en 5.2.1), y ninguno de ellos llega al "límite teórico", dado por la varianza de las pruebas subjetivas. A su vez, cada uno de estos modelos se basa en características preceptuales muy diferentes, y la capacidad de proceso requerida para implementarlos es notoriamente diferente entre sí. El hecho de que la mejora global obtenida frente al PSNR no es mucha sumada a la gran complejidad de implementación de alguno de los modelos propuestos, pone en duda su aplicabilidad.

El VQEG está trabajando en la comparación de modelos que apliquen a otros escenarios, como HDTV y Multimedia, en modelos FR, RR y NR, así como también en TV para modelos RR y NR. Se espera que en los próximos años se disponga de comparaciones similares a las realizadas para FRTV para estas aplicaciones.

Un área de particular interés es la transmisión de video y multimedia sobre redes IP, y la evaluación de la calidad perceptual objetiva en base a medidas de los parámetros de desempeño de la red. En audio, por ejemplo, se dispone del "Modelo E", estandarizado en la recomendación ITU-T G.107, el que permite predecir en forma objetiva, y dentro de ciertos márgenes de error, la calidad percibida en base a los codecs utilizados y las características de la red IP (por ejemplo, demoras, jitter, pérdida de paquetes). No existe, hasta el momento, un estándar similar para video, y hay aquí un camino de investigación.

Finalmente, en la búsqueda de un modelo perceptual, será necesario tener en cuenta no solamente la calidad propia del video, sino también otros factores que inciden directa o indirectamente en la experiencia del usuario. Por ejemplo, los estándares de medida de calidad de video no tienen en cuenta la calidad de audio asociado, siendo éste un aspecto que claramente incide en la experiencia del usuario y por lo tanto en la calidad perceptual general del contenido. Adicionalmente, el sincronismo entre el video y el audio ("lisp synchronization") juega también un papel importante en la percepción. Pequeños tiempos de defasaje entre ambas señales producen sensaciones molestas, y afectan notoriamente a la calidad percibida, aún cuando la calidad de las señales de audio y de video que se estén recibiendo sean aceptables. Por otro lado, cada aplicación tiene sus características que afectan a la experiencia general del usuario. En TV, las demoras entre el cambio de canales ("zapping") juega un papel importante. En video telefonía, la calidad conversacional (determinada fuertemente por las demoras totales punta a punta) es sumamente relevante en la calidad percibida. En HDTV la calidad de la imagen es mucho más importante que en aplicaciones pensadas para menores resoluciones (por ejemplo pantallas de PDAs).

Muchos aspectos referentes a la calidad perceptual de video están aún en proceso de comprensión, y requieren de mayores investigaciones. Se han presentado varias posibles líneas de investigación en el área, las que podrían ser parte de un trabajo de Tesis Doctoral.

En suma, se nota un interés creciente en el estudio de la calidad perceptual de video, y existen aún muchos aspectos que requieren investigación, estudios más profundos y comparaciones sistemáticas.

9. Glosario

ACR	Absolute Category Rating
AVC	Advanced Video Coding
BPOCS	Block Projection Onto Convex Sets
CIF	Common Intermediate Format (352 x 288)
CODEC	Codificador / Decodificador
CSF	Contrast Sensitivity Function
DCR	Degradation Category Rating

DCT	Discrete Cosine Transform
DSCQS	Double Stimulus Continuos Quality Scale
DSIS	Double Stimulus Impairment Scale
DWT	Discrete Wavelet Transform
ES	Elementary Streams
FR	Full Reference
GOP	Group of Pictures
HDTV	High Definition TV
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Standards Organization
ITU	International Telecommunication Union
JPEG	Joint Photographic Expert Group
LAN	Local Area Network
MC	Motion Compensation
MM	Multimedia
MOS	Mean Opinion Score
MPEG	Moving Pictures Expert Group
MSE	Mean Square Error)
MTS	MPEG Transport Stream
MVO	Multiple Video Objects
NR	No Reference
NTSC	National Television Systems Committee
PAL	Phase Alternating Line
PES	Packetized Elementary Streams
PLR	Packet Loss Rate
PSNR	Peak Signal to Noise Ratio
QFIC	Quarter Common Intermediate Format (174 x 144)
RFC	Request for Comment
RMSE	Root Mean Square Error
RR	Reduced Reference
RTCP	Real-Time Transport Control Protocol
RTP	Real-Time Transport Protocol
SD	Standard Definition (720 x 576)
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation
SSCQE	Single Stimulus Continuous Quality Evaluation
VCEG	Video Coding Expert Group
VO	Video Object
VoIP	Voz sobre IP
VOP	Video Object Plane
VQEG	Video Quality Experts Group
VQM	Video Quality Metric
WAN	Wide Area Network

10. Referencias

[1] Perceptual Quality Measurement and Control: Definition, Application and Performance
AR Prasad, R Esmailzadeh, S Winkler, T Ihara, B
4th International Symposium on Wireless Personal Multimedia Communications, Aalborg, Denmark, 2001

[2] Discrete Cosine Transform
N Ahmed, T Natrajan, K.R. Rao
IEEE Trans. Comput. Vol C-23, No 1, pp90-93, Dec 1984

[3] Trends and Perspectives in Image and Video Coding
T Sikora
IEEE Proceedings, Vol 93, No 1, January 2005

[4] ISO/IEC IS 10918-1, ITU-T Recommendation T.81 Digital compression and coding of continuous-tone still images: Requirements and guidelines, 1994

[5] ISO/IEC 15444-1:2004. JPEG2000 Image Coding System: Core coding system

[6] ISO/IEC 11172-2:1993. Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s

[7] ISO/IEC 13818-2:2000. Information technology – generic coding of moving pictures and associated audio information: Video.

[8] Digital television fundamentals: design and installation of video and audio systems
Michel Robin, Michel Poulin
ISBN 0-07-053168-4, 1998, McGraw-Hill

[9] ISO/IEC 14496-2:2001. Information technology – Coding of audio-visual objects – Part 2: Visual

[10] The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extension
Gary J. Sullivan, Pankaj Topiwala, and Ajay Luthra
SPIE Conference on Applications of Digital Image Processing XXVII
Special Session on Advances in the New Emerging Standard: H.264/AVC, August, 2004

[11] Overview of the H.264 / AVC Video Coding Standard
Thomas Wiegand, Gary J. Sullivan, Gisle Bjontegaard, and Ajay Luthra
IEEE Transactions on Circuits and Systems For Video Technology, Vol 13, July 2003

[12] Report of The Formal Verification Tests on AVC (ISO/IEC 14496-10 | ITU-T Rec. H.264)
ISO/IEC JTC1/SC29/WG11, MPEG2003/N6231
December 2003

[13] RFC 2250 Payload Format for MPEG1/MPEG2 Video
D. Hoffman et al, January 1998

[14] RFC 3016 RTP Payload Format for MPEG-4 Audio/Visual Streams
Y. Kikuchi et al, November 2000

[15] RFC 3640 RTP Payload Format for Transport of MPEG-4 Elementary Streams
J. van der Meer et al, November 2003

[16] RFC 3550: “RTP: A Transport Protocol for Real-Time Applications”, H. Schulzrinne et al (July 2003)

[17] RFC 3551: “RTP Profile for Audio and Video Conferences with Minimal Control”, H. Schulzrinne et al (July 2003)

[18] Video coding with H.264/AVC: Tools, Performance, and Complexity

Jörn Ostermann, Jan Bormans, Peter List, Detlev Marpe, Matthias Narroschke, Fernando Pereira, Thomas Stockhammer, and Thomas Wedi
IEEE Circuits and Systems Magazine, First Quarter 2004

[19] Recommendation ITU-R BT.500-11
Methodology for the subjective assessment of the quality of television pictures
06/2002

[20] Recommendation ITU-T P.910
Subjective video quality assessment methods for multimedia applications
09/1999

[21] Effect of Monitor Size on User-Level QoS of Audio-Video Transmission over IP Networks in Ubiquitous Environments
Y. Ito and S. Tasaka
IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2005, September 2005, Vol 3, pp 1806-1812

[22] Video Quality Experts Group
http://www.its.bldrdoc.gov/vqeg/

[23] DCT Quantization Noise in Compressed Images
Mark A. Robertson and Robert L. Stevenson
IEEE Transactions on Circuits and Systems For Video Technology, Vol. 15, No. 1, January 2005

[24] Digital Video Image Quality and Perceptual Coding
H.R. Wu and K.R Rao
2006, CRC Press

[25] Blocking artifact detection and reduction in compressed data
Triantafyllidis, G.A. Tzovaras, D. Strintzis, M.G.
IEEE Transaction on Circuits and Systems for Video Technology, Vol. 12, No. 10, October 2002

[26] Adaptive reduction of blocking artifacts in DCT domain for highly compressed images
Ci Wang, Wen-Jun Zhang, Xiang-Zhong Fang
IEEE Transaction on Consumer Electronics, May 2004, Vol 50. Issue 2, pp 647-654

[27] De-Blocking Artifacts in DCT Domain Using Projection onto Convex Sets Algorithm
Hai-Feng XU1, Song-Yu YU1 and Ci WANG2
IEICE Transactions on Information and Systems 2006, Vol E89-D, pp 2460-2463

[28] Blocking artifacts suppression in block-coded images using overcomplete wavelet representation
Alan W.-C. Liew, Hong Yan,
IEEE Transaction on Circuits and Systems for Video Technology, Vol 14, No 4, April 2004

[29] Multiframe blocking-artifact reduction for transform-coded video
Gunturk, B.K. Altunbasak, Y. Mersereau, R.M.
IEEE Transactions on Circuits and Systems for Video Technology, April 2002, Vol 12, Issue 4, pp 276-282

[30] Blur determination in the compressed domain using DCT information

Marichal, X. Wei-Ying Ma HongJiang Zhang
IEEE International conference on Image Processing ICIP 1999, Vol 2, pp 386-390

[31] Blur detection for digital images using wavelet transform
Hanghang Tong Mingjing Li Hongjiang Zhang Changshui Zhang
ICME IEEE International Conference on Multimedia, June 2004

[32] An adaptive postprocessing technique for the reduction of color bleeding in DCT-coded images
François-Xavier Coudoux, Marc Gazalet, Patrick Corlay
IEEE Transactions on circuits and Systems for Video Technology, Vol. 14, No. 1, January 2004

[33] Reduction of color bleeding for 4:1:1 compressed video
François-Xavier Coudoux, Marc Gazalet, Member, IEEE, and Patrick Corlay
IEEE Transaction on Broadcasting, Vol. 51, No 4. December 2005

[34] Two-stage false contour detection using directional contrast and its application to adaptive false contour reduction
Ji Won Lee, Bo Ra Lim, Rae-Hong Park, Jae-Seung Kim and Wonseok Ahn
IEEE Transactions on Consumer Electronics, Vol. 52, No. 1, pp 179-188, February 2006

[35] Comparison of blocking and blurring metrics for video compression
Leontaris, A. Reibman, A.R.
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2005

[36] Teoría de Los Sistemas: El Ojo Humano
Bruno Nicolás Gutiérrez Perea
Monografía Examen Final, Profesorado De Tecnología, 2006

[37] The handbook of Video Databases: Dessign and Applications, Chapter 41
B. Furth and O, Marqure
September 2003

[38] Digital Video Quality, Vision Models and Metrics
Stefan Winkler
John Wiley & Sons Ltd, 2005

[39] FINAL REPORT FROM THE VIDEO QUALITY EXPERTS GROUP ON THE VALIDATION OF OBJECTIVE MODELS OF VIDEO QUALITY ASSESSMENT
June, 2000

[40] FINAL REPORT FROM THE VIDEO QUALITY EXPERTS GROUP ON THE VALIDATION OF OBJECTIVE MODELS OF VIDEO QUALITY ASSESSMENT, PHASE II ©2003 VQEG
August 25, 2003

[41] RECOMENDACIÓN UIT-R BT.1683 - Técnicas de medición objetiva de la calidad de vídeo perceptual para la radiodifusión de televisión digital de definición convencional en presencia de una referencia completa
Junio 2004

[42] A computational approach to edge detection

Canny, J.
IEEE Trans. Pattern Analysis and Machine Intelligence, 1986, Vol. 8(6), p. 679-698.

[43] Morphological image segmentation applied to video quality assessment
De Alengar Lotufo, R Da Silva, W D F Falcao, A X Pessoa
IEEE Proceedings in Computer Graphics, Image Processing and Vision, SIGGRAPI Proceedings, pp 468-475
October 1998

[44] A new standardized method for objectively measuring video quality
M. H. Pinson, S. Wolf
IEEE Transactions on Broadcasting, September 2004, ,Vol 50, issue 3, pp 312-322

[45] RRNR-TV Group TEST PLAN, version 1.9
Edited at Tokyo meeting to update schedule, 28/9/2006

[46] Multimedia Group TEST PLAN, Draft Version 1.16
February 7, 2007

[47] Why is image quality assessment so difficult?
Zhou Wang Bovik, A.C. Ligang Lu
IEEE International Conference on Acoustics, Speech and Signal Processiong, 2002. Proceedings

[48] Image Quality Assessment: From error measurement to structural similarity
Zhou Wang Bovik, A.C. H.R. Sheick
IEEE Transaction on Image Processing, April 2004

[49] No-reference quality metric for degraded and enhanced video
J E Caviedes, F Oberti
Proceedings of SPIE , vol 5150, pp 621-632, 2003

[50] Video Quality Evaluation for Internet Streaming Applications
Stefan Winkler and Ruth Campos
Proceedings of SPIE, vol 5007, pp 104-115, 2003

[51] A generalized block-edge impairment metric for video coding
H. R. Wu, M. Yuen
IEEE Signal Processing Letters 4(11):317–320, 1997.

[52] Detection of blocking artifacts in compressed video
T. Vlachos
Electronics Letters 36(13):1106–1108, 2000.

[53] Blind measurement of blocking artifacts in images
Z. Wang, A. C. Bovik, B. L. Evans:
Proc. ICIP, vol. 3, pp. 981–984, Vancouver, Canada, 2000.

[54] Perceptual Video Quality and Blockiness Metrics for Multimedia Streaming Applications
Stefan Winkler, Animesh Sharma, David McNally
Proc International Symposium on Wireless Personal Multimedia Communication, 2001, pp 553-556

[55] User-Oriented QoS Analysis in MPEG-2 Video Delivery

O Verscheure, P Frossard, M Hamdi
Real Time Imaging 5, 1999, pp 305-314

[56] Quality Monitoring of Video Over a Packet Network
Amy R. Reibman, Vinay A. Vaishampayan and Yegnaswamy Sermadevi
IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 6, NO. 2, APRIL 2004

[57] Visibility of individual packet losses in MPEG-2 video
Amy R. Reibman, Sandeep Kanumuri, Vinay Vaishampayan and Pamela C. Cosman
IEEE International Conference on Image Procecssing) 2004, ICIP'04, Vol 1, pp 171-174

[58] Delivery of MPEG Video Streams with constant perceptual quality of service
Quaglia, D. De Martin, J.C.
IEEE, Proceedings International Conference on Multimedia, 2002

[59] Estimation of packet loss effects on video quality
Bouazizi, I.
IEEE, First International Symposium on Control, Communications and Signal Processing, 2004, pp 91-94.

[60] Packet Loss Resilience for MPEG-4 Video Stream over the Internet
Jae-Young Pyun, Jae-Han Jung, Jae-Jeong Shim
IEEE Transactions on consumer electronics, Vol 48, issue 3, August 2002

[61] Adaptive Media Playout of Low Delay Video Streaming Over Error Prone Channels
Mark Kalman, Eckehard Steinbach, Bernd Girod,
IEEE Transactions on Circuits and Systems for Video Technology, Vol 14, no 6, June 2004

[62] Recommendation ITU-T G.107
The E-model, a computational model for use in transmission planning
March 2005

[63] TIA/TSB 116-A Telecommunications - IP Telephony Equipment – Voice Quality Recommendations for IP Telephony
March 1, 2006

[64] A Passive Method for Monitoring Voice-over-IP Call Quality with ITU-T Objective Speech Quality Measurement Methods
Conway, A.E.
IEEE International Conference on Communications, ICC 2002, Vol 4, pp 2583-2586