UNIÓN INTERNACIONAL DE TELECOMUNICACIONES

UIT-T J.144

(03/2001)

SECTOR DE NORMALIZACIÓN
DE LAS TELECOMUNICACIONES
DE LA UIT

SERIE J: REDES DE CABLE Y TRANSMISIÓN DE PROGRAMAS RADIOFÓNICOS Y TELEVISIVOS, Y DE OTRAS SEÑALES MULTIMEDIOS

Mediciones de la calidad de servicio

Técnicas de medición objetiva de la percepción de la calidad vídeo en televisión por cable en presencia de una referencia completa

Recomendación UIT-T J.144

(Anteriormente Recomendación del CCITT)

RECOMENDACIONES UIT-T DE LA SERIE J

REDES DE CABLE Y TRANSMISIÓN DE PROGRAMAS RADIOFÓNICOS Y TELEVISIVOS, Y DE OTRAS SEÑALES MULTIMEDIOS

Recomendaciones generales	J.1–J.9
Especificaciones generales para transmisiones radiofónicas analógicas	J.10–J.19
Características de funcionamiento de los circuitos radiofónicos	J.20–J.29
Equipos y líneas utilizados para circuitos radiofónicos analógicos	J.30–J.39
Codificadores digitales para señales radiofónicas analógicas	J.40–J.49
Transmisión digital de señales radiofónicas	J.50–J.59
Circuitos para transmisiones de televisión analógica	J.60–J.69
Transmisiones de televisión analógica por líneas metálicas e interconexión con radioenlaces	J.70–J.79
Transmisión digital de señales de televisión	J.80–J.89
Servicios digitales auxiliares para transmisiones de televisión	J.90–J.99
Requisitos operacionales y métodos para transmisiones de televisión	J.100–J.109
Sistemas interactivos para distribución de televisión digital	J.110–J.129
Transporte de señales MPEG-2 por redes de transmisión de paquetes	J.130–J.139
Mediciones de la calidad de servicio	**J.140–J.149**
Distribución de televisión digital por redes locales de abonados	J.150–J.159
IPCablecom	J.160–J.179
Varios	J.180–J.199
Aplicación para televisión digital interactiva	J.200–J.209

Para más información, véase la Lista de Recomendaciones del UIT-T.

Recomendación UIT-T J.144

Técnicas de medición objetiva de la percepción de la calidad vídeo en televisión por cable en presencia de una referencia completa

Resumen

Esta Recomendación contiene directrices sobre la selección del equipo apropiado de medición objetiva de la percepción de la calidad vídeo, diseñado para su utilización en aplicaciones de televisión por cable digital cuando se dispone de la señal de vídeo de referencia completa.

Orígenes

La Recomendación UIT-T J.144, preparada por la Comisión de Estudio 9 (2001-2004) del UIT-T, fue aprobada por el procedimiento de la Resolución 1 de la AMNT el 9 de marzo de 2001.

PREFACIO

La UIT (Unión Internacional de Telecomunicaciones) es el organismo especializado de las Naciones Unidas en el campo de las telecomunicaciones. El UIT-T (Sector de Normalización de las Telecomunicaciones de la UIT) es un órgano permanente de la UIT. Este órgano estudia los aspectos técnicos, de explotación y tarifarios y publica Recomendaciones sobre los mismos, con miras a la normalización de las telecomunicaciones en el plano mundial.

La Asamblea Mundial de Normalización de las Telecomunicaciones (AMNT), que se celebra cada cuatro años, establece los temas que han de estudiar las Comisiones de Estudio del UIT-T, que a su vez producen Recomendaciones sobre dichos temas.

La aprobación de Recomendaciones por los Miembros del UIT-T es el objeto del procedimiento establecido en la Resolución 1 de la AMNT.

En ciertos sectores de la tecnología de la información que corresponden a la esfera de competencia del UIT-T, se preparan las normas necesarias en colaboración con la ISO y la CEI.

NOTA

En esta Recomendación, la expresión "Administración" se utiliza para designar, en forma abreviada, tanto una administración de telecomunicaciones como una empresa de explotación reconocida de telecomunicaciones.

PROPIEDAD INTELECTUAL

La UIT señala a la atención la posibilidad de que la utilización o aplicación de la presente Recomendación suponga el empleo de un derecho de propiedad intelectual reivindicado. La UIT no adopta ninguna posición en cuanto a la demostración, validez o aplicabilidad de los derechos de propiedad intelectual reivindicados, ya sea por los miembros de la UIT o por terceros ajenos al proceso de elaboración de Recomendaciones.

En la fecha de aprobación de la presente Recomendación, la UIT no ha recibido notificación de propiedad intelectual, protegida por patente, que puede ser necesaria para aplicar esta Recomendación. Sin embargo, debe señalarse a los usuarios que puede que esta información no se encuentre totalmente actualizada al respecto, por lo que se les insta encarecidamente a consultar la base de datos sobre patentes de la TSB.

ÍNDICE

Introducción

La televisión digital da lugar a nuevas consideraciones con respecto a la calidad de servicio, con complejas relaciones entre las mediciones objetivas de los parámetros y la calidad subjetiva de la imagen. Si bien es conveniente efectuar las mediciones objetivas con una buena correlación con la evaluación subjetiva de la calidad para obtener una calidad de servicio óptima en el funcionamiento de los sistemas de televisión por cable, cabe recordar que las mediciones objetivas no sustituyen directamente la evaluación subjetiva de la calidad.

Las evaluaciones subjetivas de la calidad son procedimientos cuidadosamente diseñados para determinar la opinión promedio de observadores humanos sobre una serie de secuencias de vídeo para una aplicación determinada. Los resultados de esas pruebas se utilizan en la preparación de sistemas básicos y evaluaciones de referencia. Las evaluaciones subjetivas de la calidad para una aplicación diferente, en condiciones de prueba distintas, seguirán proporcionando resultados significativos, aunque probablemente difieran las notas de opinión sobre la misma serie de secuencias de vídeo. Las mediciones objetivas se utilizan en una amplia gama de aplicaciones produciendo los mismos resultados con un conjunto determinado de secuencias de vídeo. La selección de las secuencias de vídeo que se han de utilizar y la interpretación de las mediciones objetivas resultantes son algunos de los factores que varían para una aplicación específica.

Las mediciones objetivas y la evaluación subjetiva de la calidad son, por tanto, complementarias pero no intercambiables. Si la evaluación subjetiva se utiliza con fines relacionados a la investigación, las mediciones objetivas se necesitan para las especificaciones de equipos y la medición y el control diarios de la calidad de funcionamiento de los sistemas.

Para los fines de esta Recomendación se han adoptado las siguientes definiciones.

– "Evaluación subjetiva", que se refiere a la determinación de la calidad o el deterioro de imágenes de tipo programa presentadas a un grupo de evaluadores humanos en sesiones de observación visual.

– "Medición perceptiva objetiva", que se refiere a la medición de las características de una cadena de programa utilizando imágenes de tipo programa y métodos de medición objetiva (con instrumentos) para obtener un indicio que se acerque a la nota que se obtendría con una prueba de evaluación subjetiva.

– "Medición de la señal", que se refiere a la medición de la calidad de una cadena de programa utilizando señales de prueba y métodos de medición objetiva (con instrumentos).

En la presente Recomendación los términos medición objetiva y medición perceptiva se utilizan indistintamente para referirse a la medición perceptiva objetiva.

Hay tres métodos básicos para efectuar las mediciones objetivas:

• FR – Método aplicable cuando se dispone de la señal de vídeo de referencia. Se trata del método de dos extremos, del cual se ocupa la presente Recomendación.

• RR – Método aplicable cuando sólo se dispone de información de referencia de vídeo reducida. Se trata también de un método de dos extremos que es objeto de una Recomendación aparte [en estudio].

• NR – Método aplicable cuando no se dispone de señal de vídeo de referencia ni de información. Se trata de un método de un solo extremo y es objeto de una Recomendación aparte [en estudio].

Cada método tiene sus propias aplicaciones, y proporcionan grados diferentes de exactitud de la medición, expresada en términos de correlación con los resultados de la evaluación subjetiva.

Recomendación UIT-T J.144

Técnicas de medición objetiva de la percepción de la calidad vídeo en televisión por cable en presencia de una referencia completa

1 Alcance

Esta Recomendación contiene directrices sobre la selección del equipo apropiado de medición de la percepción de la calidad vídeo, diseñado para su utilización en aplicaciones de televisión por cable digital cuando se puede utilizar el método de medición con referencia completa.

La utilización prevista del método de medición con referencia completa es en aquellas situaciones en que se dispone fácilmente de una señal de vídeo de referencia no degradada en el punto de medición, como puede ser el caso de las mediciones en equipos individuales o en una cadena en el laboratorio o en un entorno cerrado, por ejemplo, el extremo de cabecera de la televisión por cable.

2 Referencias

2.1 Referencias normativas

Las siguientes Recomendaciones del UIT-T y otras referencias contienen disposiciones que, mediante su referencia en este texto, constituyen disposiciones de la presente Recomendación. Al efectuar esta publicación, estaban en vigor las ediciones indicadas. Todas las Recomendaciones y otras referencias son objeto de revisiones por lo que se preconiza que los usuarios de esta Recomendación investiguen la posibilidad de aplicar las ediciones más recientes de las Recomendaciones y otras referencias citadas a continuación. Se publica periódicamente una lista de las Recomendaciones UIT-T actualmente vigentes.

– UIT-R BT.500-9 (1998), *Metodología para la evaluación subjetiva de la calidad de las imágenes de televisión*.

2.2 Referencias informativas

– UIT-T J.140 (1998), *Evaluación subjetiva de la calidad de imagen para sistemas de televisión digital por cable*.

– UIT-T J.143 (2000), *Requisitos de usuario para mediciones objetivas de la percepción de la calidad vídeo en televisión digital por cable*.

– UIT-T P.910 (1996), *Métodos de evaluación subjetiva de la calidad vídeo para aplicaciones multimedios*.

– Contribución COM 9-80 de la Comisión de Estudio 9 del UIT-T (2000), *Informe Final del Grupo de Expertos en calidad vídeo sobre la validación de modelos objetivos de evaluación de la calidad vídeo.*

3 Términos, definiciones y acrónimos

En esta Recomendación se definen los términos siguientes.

3.1 evaluación subjetiva: Determinación de la calidad o el deterioro de imágenes de tipo programa presentadas a un grupo de evaluadores humanos en sesiones de observación visual.

3.2 medición perceptiva objetiva: Medición de las características de una cadena de programa utilizando imágenes de tipo programa y métodos de medición objetiva (con instrumentos) para obtener un indicio que se acerque a la nota que se obtendría con una prueba de evaluación subjetiva.

3.3 medición de la señal: Medición de la calidad de una cadena de programa utilizando señales de prueba y métodos de medición objetiva (con instrumentos).

4 Requisitos de usuario

En UIT-T J.143 se dan los requisitos de usuario para la medición objetiva de la percepción de la calidad de imagen.

5 Descripción del método de medición con referencia completa

El método de medición de dos extremos con referencia completa, para la medición objetiva de la calidad de vídeo percibida, evalúa la calidad de funcionamiento de los sistemas estableciendo una comparación entre la entrada no distorsionada, o referencia, que es la señal de vídeo a la entrada del sistema, y la señal degradada a la salida del sistema (figura 1).

La figura 1 muestra un ejemplo de aplicación del método con referencia total para probar un códec en laboratorio.

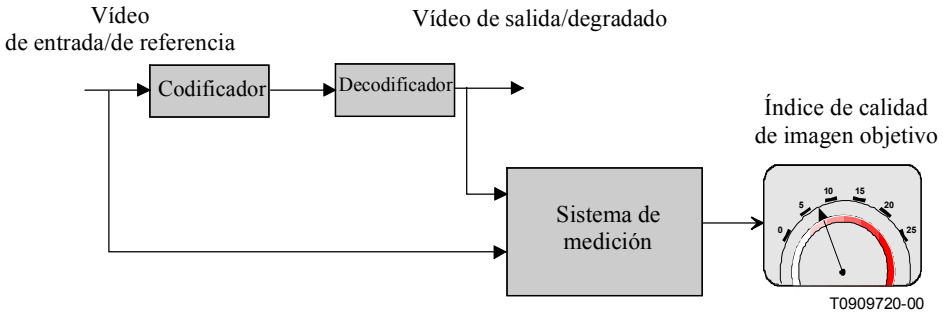


Figura 1/J.144 – Aplicación del método de medición de la calidad percibida con referencia completa para la prueba de un códec en laboratorio

La comparación entre las señales de entrada y salida requiere un proceso de alineación espacial y temporal para compensar cualquier arrastre o desplazamiento vertical u horizontal de la imagen. Quizás también se requiera corregir cualesquiera desplazamientos o diferencias de ganancia en los canales de luminancia y crominancia. A continuación se calcula el índice objetivo de la calidad de la imagen, aplicando normalmente un modelo perceptivo de la visión humana.

Puesto que la herramienta utilizada para efectuar el diagnóstico se basa en un modelo de la visión humana, más bien que en la medición de los efectos visuales secundarios específicos de la codificación, en principio vale igualmente para sistemas analógicos y sistemas digitales. También en principio es válido para cadenas en las que se combinan sistemas analógicos y digitales, o cuando se concatenan sistemas de compresión digital.

La figura 2 muestra un ejemplo de la aplicación del método con referencia completa para la prueba de una cadena de transmisión.

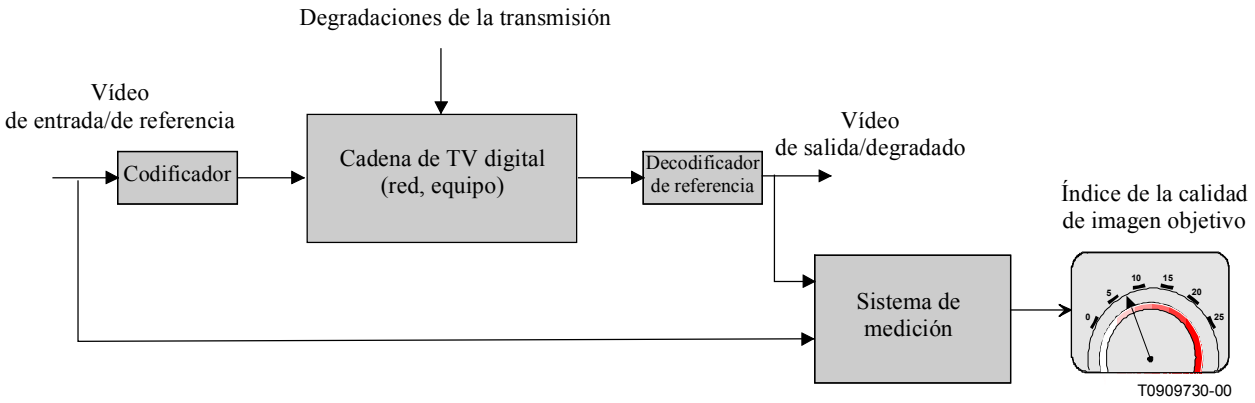


Figura 2/J.144 – Aplicación del método de medición de la calidad percibida con referencia completa para la prueba de una cadena de transmisión

En este caso, la entrada al decodificador de referencia procede de diversos puntos de la cadena de transmisión; el decodificador puede estar situado, por ejemplo, en un punto de la red, como en la figura 2, o directamente a la salida del codificador como en la figura 1. Si la cadena de transmisión digital es transparente, la medición del índice de la calidad de imagen objetivo en la fuente es igual a la medición en cualquier punto subsiguiente de la cadena.

Por lo general, se acepta que con el método de referencia completa se obtiene la mayor exactitud de las mediciones de la calidad de imagen percibida. El método ha demostrado la posibilidad de establecer un alto grado de correlación con las evaluaciones subjetivas efectuadas de conformidad con la especificación de la escala de calidad continua de doble estímulo (DSCQS) de UIT-R BT.500.

6 Conclusiones del Grupo de expertos en calidad vídeo (VQEG, *video quality experts group*)

Los estudios sobre la medición de la calidad vídeo percibida se llevan a cabo en un grupo oficioso, llamado Grupo de expertos en calidad vídeo (VQEG), que informa de sus resultados a las Comisiones de Estudio 12 y 9 del UIT-T y a la Comisión de Estudio 6 del UIT-R. Lo primero que hizo el VQEG fue evaluar el rendimiento de los algoritmos propuestos para la medición de la calidad vídeo percibida con dos extremos.

El VQEG publicó un extenso proyecto final de informe sobre la primera fase de su trabajo en marzo de 2000.

Se aconseja a los lectores que estudien ese informe para adquirir un conocimiento profundo del trabajo efectuado por el VQEG hasta esa fecha. En pocas palabras, el informe expone los resultados de las pruebas efectuadas con 10 modelos presentados al VQEG por 10 proponentes diferentes, utilizados para calcular notas objetivas en comparación con la evaluación subjetiva efectuada en una amplia gama de sistemas vídeo y secuencias fuente. En las pruebas se contrastó el comportamiento de los modelos de los proponentes con las pruebas de evaluación subjetiva de las mismas imágenes, y también con al algoritmo de "referencia" relación señal/ruido de cresta (PSNR, *peak signal-to-noise ratio*). Lo que se pretendía era comprobar los modelos propuestos en términos de:

- exactitud de la predicción (la capacidad del modelo de predecir la calidad subjetiva);
- monotonicidad de la predicción (el grado en que las predicciones del modelo concuerdan con la clasificación por categorías de los índices de calidad subjetivos);
- coherencia de la predicción (el grado en que el modelo mantiene la exactitud de la predicción con diversas secuencias de prueba de vídeo y sistemas de vídeo, es decir, en qué medida su respuesta se mantiene con diversas degradaciones de vídeo).

Se generaron más de 26 000 notas de opinión subjetivas en base a 20 secuencias fuente diferentes procesadas por 16 sistemas de vídeo diferentes y evaluadas en ocho laboratorios independientes de todo el mundo.

Las pruebas subjetivas se clasificaron en cuatro categorías: 50 Hz/alta calidad, 50 Hz/baja calidad, 60 Hz/alta calidad y 60 Hz/baja calidad. Alta calidad en este contexto se refiere a la calidad de la producción de vídeo y baja calidad se refiere a la calidad de la distribución de vídeo. Las categorías de alta calidad comprendían vídeo a velocidades binarias entre 3 Mbit/s y 50 Mbit/s. Las categorías de baja calidad comprendían vídeo a velocidades binarias entre 768 kbit/s y 4,5 Mbit/s.

En la evaluación subjetiva se siguieron de manera estricta los procedimientos de UIT-R BT.500-9 para el método de escala de calidad continua de doble estímulo (DSCQS, *double stimulus continuous quality scale*). Los planes de pruebas subjetivas y objetivas incluían procedimientos de análisis de validación de las notas subjetivas y cuatro medidas para comparar los datos objetivos con los resultados subjetivos.

Además del análisis basado en el conjunto de datos totales, se efectuaron análisis de subconjuntos basados en las cuatro categorías de pruebas subjetivas y los datos totales, con exclusión de ciertos sistemas de procesamiento de vídeo, para determinar la correlación entre los resultados y diversos parámetros dependientes de la aplicación.

Se descartaron los resultados obtenidos con los dos algoritmos que no se probaron totalmente o que se encontró que tenían problemas de aplicación práctica. Los resultados de las pruebas del VQEG, basados en el análisis obtenido en las cuatro categorías de pruebas subjetivas, muestran básicamente lo siguiente:

- Ningún sistema de medición objetiva en la prueba es capaz de sustituir por completo las pruebas subjetivas.
- Ningún modelo objetivo supera estadísticamente a los otros modelos en todas las condiciones de referencia.
- Ningún modelo objetivo supera estadísticamente la PSNR en todas las condiciones de referencia.
- En base a los datos de que se dispone actualmente, no cabe recomendar por ahora a la UIT ningún método.

En el lado positivo, el trabajo efectuado por el VQEG ha dado lugar a una comprensión mucho mayor del problema de la prueba de la calidad de vídeo percibida, y de los requisitos de los usuarios. Es probable que aquí se derive el desarrollo de modelos perceptivos mejorados, implementados en equipos comerciales.

En el Subcomité G-2.1.6 del IEEE se ha planificado la realización de estudios para proporcionar un conjunto de escenas de prueba degradadas de manera controlada. Cada escena tendrá su correspondiente escala de percepción, calibrada en pasos sucesivos de diferencias apenas perceptibles de degradación. Se espera que dichas escenas constituyan un buen fondo de material de referencia con el que probar los sistemas del futuro.

7 Conclusiones

Puesto que por ahora no se puede recomendar ningún método de medición, en esta cláusula se hacen algunas recomendaciones de carácter general sobre los modelos de evaluación de la calidad vídeo que utilizan la metodología de la referencia completa. En los apéndices se detallan los modelos actuales validados por el VQEG. Está previsto adoptar como norma uno o más de esos modelos (o nuevos modelos que se propongan), en base al trabajo futuro del VQEG y de otros grupos. La futura labor del VQEG probablemente incluya además otras condiciones de prueba, tales como, por ejemplo, distancias de visión más cercanas y tipos y gamas adicionales de distorsión, lo cual permitirá diferenciar mejor entre los modelos objetivos y entre cada modelo y la PSNR.

Recomendación general

Cuando se efectúen mediciones de la calidad vídeo percibida, utilizando el modelo con referencia completa descrito en esta Recomendación, los operadores deberán analizar primero la equivalencia de sus requisitos de aplicación y usuario específicos en términos de características y funcionamiento del equipo de medición.

A continuación se indican algunos aspectos que se han de tomar en consideración:

- coste imputable a la propiedad del equipo de medición perceptiva;
- servicio posventa del fabricante;
- facilidad de funcionamiento;
- fiabilidad;
- tamaño, peso y necesidad de alimentación de potencia;
- rapidez de la medición en tiempo real y en tiempo no real;
- funcionamiento en línea (en servicio);
- exactitud, monotonicidad y coherencia de la predicción.

Cuando se notifiquen los resultados de las mediciones de la calidad vídeo percibida, los operadores deberán indicar siempre la marca, el modelo y los ajustes del equipo de medición perceptiva y las imágenes de prueba utilizadas. De esta manera, los operadores podrán comparar los resultados de esas pruebas con los de las pruebas efectuadas por otros operadores.

Esa precaución es necesaria porque quizás el equipo de medición perceptiva con referencia total proporcione un grado de correlación con las pruebas de medición subjetiva que dependa, entre otros factores, del conjunto de imágenes de prueba seleccionados, del grado de compresión aplicada al tren binario de vídeo objeto de la prueba y del número de posibles implementaciones que el fabricante haya tenido en cuenta en su diseño.

Cuando los operadores duden respecto a la elección del equipo de medición perceptiva con referencia total de entre los modelos disponibles en el mercado, o antes de que decidan optar por un modelo nuevo, harían bien en efectuar un conjunto de pruebas con el equipo nuevo, comprobando la correlación de sus indicaciones con las obtenidas mediante pruebas de medición subjetivas efectuadas con un conjunto apropiado de imágenes o secuencias de prueba.

Modelos de calidad vídeo objetiva – El camino hacia futuras revisiones

Por último, como ayuda en la selección del modelo de medición perceptiva que mejor se atiene a sus necesidades, los operadores pueden consultar el apéndice I. Dicho apéndice se basa en el informe final del VQEG que figura en la Contribución COM 9-80 a la Comisión de Estudio 9 del UIT-T, junio de 2000.

El apéndice I va a ser actualizado periódicamente para reflejar el trabajo en curso del VQEG y otros grupos, así como la experiencia práctica que los participantes del trabajo UIT-T puedan adquirir utilizando los equipos de medición perceptiva.

A medida que mejoren los métodos del apéndice I (u otros que puedan proponerse más adelante), se divulguen plenamente y adquieran una mayor validez, podrán ser adoptados como secciones normativas de la presente Recomendación. Para que un modelo se convierta en norma, debe ser comprobado por un órgano abierto independiente (como el VQEG) que efectúe la evaluación técnica en base a las directrices y los criterios de calidad establecidos por la Comisión de Estudio 9 (CE 9). Lo que pretende a la postre la CE 9 es recomendar un solo método normativo de referencia completa para la televisión por cable.

APÉNDICE I

Modelos de medición de la percepción de la calidad vídeo con referencia completa

I.1 Descripción de los modelos

En los apéndices I a IX se describen los ocho modelos que fueron validados por el Grupo de expertos en calidad vídeo (VQEG) y documentados en el informe final de dicho Grupo de marzo de 2000. A continuación se indican algunos detalles de esos modelos junto con una descripción de la relación señal/ruido de cresta (PSNR).

I.1.1 PSNR

La PSNR se define de acuerdo con las fórmulas siguientes:

$$PSNR = 10 \log_{10}\left(\frac{255^2}{MSE}\right)$$

$$MSE = \frac{1}{(P2-P1+1)(M2-M1+1)(N2-N1+1)} \sum_{p=P1}^{p=P2} \sum_{m=M1}^{m=M2} \sum_{n=N1}^{n=N2} (d(p,m,n)-o(p,m,n))^2$$

donde *d(p,m,n)* y *o(p,m,n)* representan respectivamente el valor del píxel degradado y original en la trama p, fila m y columna n.

NOTA – La PSNR requiere el empleo, con seguridad, de un alto grado de normalización. Para la normalización hace falta efectuar alineaciones espaciales y temporales así como correcciones por ganancias y desplazamientos. El método de normalización es objeto de otra Recomendación (en estudio).

I.1.2 CPqD

Al modelo de CPqD presentado para las pruebas del VQEG se le ha dado la denominación temporal de CPqD-IES (Evaluación de la imagen basada en la segmentación, *image evaluation based on segmentation*) versión 2.0. La primera versión de este sistema de evaluación objetiva de la calidad, el CPqD-IES versión 1.0, era un sistema concebido para poder hacer predicciones de la calidad con un conjunto de escenas predefinidas.

El CPqD-IES versión 1.0 implementa la evaluación de la calidad vídeo utilizando parámetros objetivos basados en la segmentación de la imagen. Escenas naturales se segmentan en regiones de plano, de borde y de textura, y se asigna un conjunto de parámetros objetivos a cada uno de esos contextos. Se define un modelo basado en la percepción que predice índices subjetivos calculando la relación entre medidas y resultados objetivos de las pruebas de evaluación subjetiva, aplicadas a un conjunto de escenas naturales procesadas mediante sistemas de procesamiento de vídeo. En este modelo, la relación entre cada parámetro objetivo y el nivel de degradación subjetivo se aproxima mediante una curva logística con la que se obtiene un nivel de degradación estimado para cada parámetro. Al resultado final se llega mediante una combinación de niveles de degradación estimados, basados en sus respectivas fiabilidades estadísticas.

Se añadió un clasificador de escenas al CPqD-IES versión 2.0 para obtener un sistema de evaluación independiente de las mismas. El clasificador utiliza información espacial [basada en el análisis de la transformada de coseno discreta (DCT)] e información temporal (basada en cambios de la segmentación) de la secuencia de entrada para obtener los parámetros del modelo a partir de una base de datos de 12 escenas (525/60 Hz) (apéndice II).

I.1.3 Tektronix/Sarnoff

La contribución de Tektronix/Sarnoff se basa en un modelo de discriminación visual que simula las respuestas de los mecanismos visuales espacio-temporales del ser humano y las magnitudes percibidas de las diferencias a la salida de los mecanismos entre secuencias fuente y secuencias

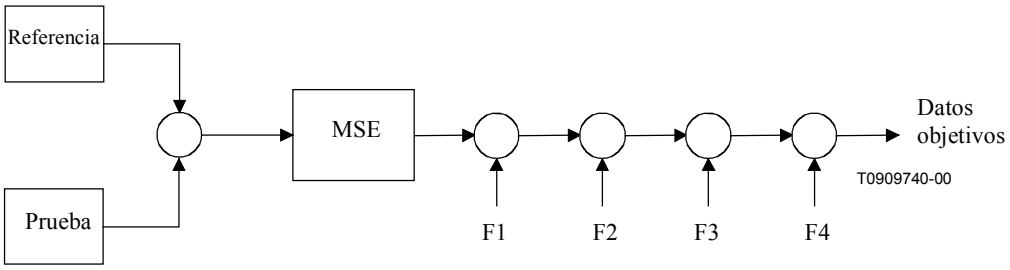
procesadas. A partir de esas diferencias se calcula una medida global de las posibilidades de distinguir entre las dos secuencias. El modelo se concibió con el requisito de funcionamiento a alta velocidad con soporte físico de tratamiento de imágenes estándar, y representa por ello una solución relativamente directa y de fácil cálculo (apéndice III).

I.1.4 NHK/Mitsubishi Electric Corp.

Este modelo emula las características visuales del ser humano utilizando filtros tridimensionales (espacio-temporales), que se aplican a las diferencias entre señales fuente y señales procesadas. Las características del filtro se varían dependiendo del nivel de luminancia. La nota de calidad resultante se calcula como la suma de las medidas ponderadas de los filtros. La versión del soporte físico ahora disponible puede medir la calidad de la imagen en tiempo real y se va a utilizar en diversos entornos de radiodifusión, tales como el de supervisión en tiempo real de las señales radiodifundidas (apéndice IV).

I.1.5 KDD

Véase la figura I.1.



F1 Filtrado espacial basado en píxels
F2 Filtrado basado en bloques (efecto del enmascaramiento del ruido)
F3 Filtrado basado en tramas (dispersión del punto de observación)
F4 Filtrado basado en secuencias (vector de movimiento + segmentación del objeto, etc.)

Figura I.1/J.144 – Descripción del modelo

El error cuadrático medio (MSE) se calcula sustrayendo la señal de prueba (Prueba) de la señal de referencia (Referencia) y se pondera con los filtros visuales humanos F1, F2, F3 y F4.

El modelo presentado es el F1+F2+F4 (versión 2.0, agosto de 1998) (apéndice V).

I.1.6 EPFL

La medida de la distorsión perceptiva (PDM, *perceptual distortion metric*) presentada por EPFL se basa en un modelo espacio-temporal del sistema visual humano. Consta de cuatro etapas, a través de las cuales pasan la secuencia de referencia y la secuencia procesada. En la primera se convierte la entrada en un espacio de colores antagónicos. La segunda etapa implementa una descomposición perceptiva espacio-temporal en canales visuales separados de frecuencia temporal, frecuencia espacial y orientación diferentes. La tercera etapa modela los efectos del enmascaramiento del diagrama simulando mecanismos estimuladores e inhibidores de acuerdo con un modelo de control de la ganancia del contraste. La cuarta y última etapa de la medida es una etapa de agrupación y detección en la que se calcula una medida de la distorsión a partir de la diferencia entre las salidas de los sensores de la secuencia de referencia y la secuencia procesada (apéndice VI).

I.1.7 NASA

El modelo propuesto por la NASA se denomina calidad vídeo digital (DVQ, *digital video quality*) y su versión es la 1.08b. Con esta medida se intenta incorporar muchos aspectos de la sensibilidad

visual humana en un algoritmo de procesamiento de imágenes sencillo. La sencillez es un objetivo importante ya que lo ideal sería que la medición se efectuara en tiempo real y sólo se necesitaran unos pocos recursos de cálculo. Uno de los elementos más complejos de las otras mediciones propuestas y al que más tiempo hay que dedicar son las operaciones de filtrado espacial efectuadas para implementar los filtros espaciales múltiples de paso de banda característicos de la visión humana. Esta etapa se acelera utilizando la transformada de coseno discreta (DCT, *discrete cosine transform*) para la descomposición en canales espaciales. Esto representa una ventaja importante ya que se dispone de soportes físico y lógico eficaces para efectuar esa transformación, y porque es posible que, en muchas aplicaciones, la transformación ya se haya realizado como parte del proceso de compresión.

Los datos que se introducen en la medición son un par de secuencias de imágenes en color: la de referencia y la de prueba. El primer paso consiste en la realización de diversos muestreos, arrastres de imágenes y transformaciones del color que sirven para limitar el procesamiento a la región que interesa y para expresar las secuencias en un espacio de percepción del color. En esta etapa se efectúa también el desentrelazado y la descorrección gamma del vídeo de entrada. Las secuencias se someten a continuación a una división en bloques y a una transformada de coseno discreta (DCT), y los resultados se convierten seguidamente en contraste local. Los pasos siguientes son los filtrados temporal y espacial, y una operación de enmascaramiento del contraste. Por último, las diferencias enmascaradas se agrupan en dimensiones espacio-temporales y cromáticas para calcular una medida de la calidad (apéndice VII).

I.1.8 KPN/Swisscom CT

La medición de la percepción de la calidad vídeo (PVQM, *perceptual video quality measure*) desarrollada por KPN/Swisscom CT utiliza el mismo procedimiento de medida de la calidad vídeo que la medición de la calidad vocal perceptual (PSQM, *perceptual speech quality measure* [1], UIT-T P.861 [2]) para medir la calidad de las señales vocales. El método se concibió con miras a evaluar las distorsiones espaciales y temporales y las de ubicación espacio-temporal como las que se producen en condiciones de error. Utiliza las secuencias de vídeo de formato de entrada de UIT-R BT.601-5 [3] (entrada y salida) y las muestrea de nuevo a 4:4:4, formato Y, Cb, Cr. En el algoritmo se incluye una alineación de la luminancia espacio-temporal. Puesto que los cambios globales de brillo y contraste repercuten sólo de manera limitada en la calidad percibida de manera subjetiva, la PVQM utiliza una adaptación de brillo/contraste especial de la secuencia de vídeo distorsionada. La alineación espacio-temporal se lleva a cabo aplicando un procedimiento que busca la concordancia de los bloques. La parte análisis de la luminancia espacial se basa en la detección del borde de la señal Y, mientras que la parte análisis temporal se basa en el análisis de las tramas de diferencia de la señal Y. Es bien conocido el hecho de que el sistema visual humano (HVS, *human visual system*) es mucho más sensible a los cambios agudos del componente luminancia que a la del componente crominancia. Además, el HVS tiene una función de sensibilidad al contraste que disminuye a frecuencias espaciales elevadas. Estas características básicas del HVS han quedado reflejadas en el primer paso del algoritmo PVQM, que proporciona un aproximación de primer orden a las funciones de sensibilidad al contraste de las señales de luminancia y crominancia. En el segundo paso se calcula la intensidad del contorno de la luminancia Y como una representación de la señal que contiene los aspectos más importantes de la imagen. La intensidad se computa calculando el gradiente local de la señal de luminancia (utilizando un filtrado espacial de tipo Sobel) en cada trama y promediando a continuación esa intensidad a lo largo del espacio y el tiempo. En el tercer paso, se calcula el error de crominancia como una media ponderada del error de color de ambos componentes Cb y Cr con predominio del componente Cr. En el último paso, los tres indicadores diferentes se convierten en un indicador de calidad único, aplicando una sencilla regresión lineal múltiple, que correlaciona adecuadamente la calidad vídeo global de la secuencia percibida de manera subjetiva (apéndice VIII).

I.1.9 NTIA

Este modelo de calidad vídeo utiliza características de anchura de banda reducida extraídas de regiones espacio-temporales (S-T, *spatial-temporal regions*) de escenas de vídeo de entrada y salida procesadas. Esas características describen el detalle espacial, el movimiento y el color presentes en la secuencia de vídeo. Las características espaciales describen la actividad de los bordes de la imagen, o los gradientes espaciales. Los sistemas de vídeo digital pueden añadir bordes (por ejemplo, ruido de borde, división en bloques) o reducir bordes (por ejemplo, borrosidad). Las características temporales describen la actividad de las diferencias temporales, o los gradientes temporales entre tramas sucesivas. Los sistemas de vídeo digital pueden añadir movimiento (por ejemplo, bloques de errores) o reducir movimiento (por ejemplo, repeticiones de tramas). Las características de la crominancia describen la actividad de la información de color. Los sistemas de vídeo digital pueden añadir información de color (por ejemplo, diafonía cromática) o reducir información de color (por ejemplo, submuestreo de colores). Los parámetros de ganancia y pérdida se calculan comparando dos trenes paralelos de muestras de características, uno procedente de la entrada y otro de la salida. Dichos parámetros se examinan separadamente para cada par de trenes de características ya que miden fundamentalmente aspectos diferentes de la percepción de la calidad. Las funciones de comparación de características utilizadas para calcular ganancias y pérdidas tratan de emular la perceptibilidad de las degradaciones modelando umbrales de perceptibilidad, el enmascaramiento visual y la agrupación de errores. Se utiliza una combinación lineal de los parámetros para estimar el índice de calidad subjetivo (apéndice IX).

I.2 Referencias

[1] BEERENDS (J.G.) y STEMERDINK (J.A.): A perceptual speech quality measure based on a psychoacoustic sound representation*, J. Audio Eng. Soc.* 42, 115-123, 1994.

[2] UIT-T P.861 (1998), *Medición objetiva de la calidad de los códecs vocales de banda telefónica (300-3400 Hz)*.

[3] UIT-R BT.601-5 (1995), *Parámetros de codificación de televisión digital para estudios con formatos de imagen normal 4:3 de pantalla ancha 16:9*.

APÉNDICE II

CPqD

Evaluación de la calidad vídeo utilizando parámetros objetivos basados en la segmentación de la imagen

Resumen

Este apéndice presenta una metodología de evaluación de la calidad vídeo utilizando parámetros objetivos basados en la segmentación de la imagen. Escenas naturales se segmentan en regiones de plano, borde y textura, y se asigna un conjunto de parámetros objetivos a cada uno de esos contextos. Se define un modelo basado en la percepción que predice los índices subjetivos calculando la relación entre mediciones objetivas y resultados de las pruebas de evaluación subjetiva aplicadas a un conjunto de escenas naturales y códecs de vídeo MPEG-2. En este modelo, la relación entre cada parámetro objetivo y el nivel de degradación subjetivo se aproxima mediante una curva logística, con la que se obtiene un nivel de degradación estimado para cada parámetro. Al resultado final se llega mediante una combinación lineal de niveles de degradación estimados, en la que la ponderación de cada nivel de degradación es proporcional a su fiabilidad estadística. Los resultados presentados en este apéndice muestran que la utilización de mediciones objetivas basadas en regiones proporciona predicciones más exactas en comparación con las predicciones basadas en parámetros globales.

II.1 Introducción

La evaluación de la calidad vídeo se ha convertido en un asunto crucial con la utilización creciente de sistemas de compresión de vídeo digital y los servicios vídeo subsiguientes, tales como la distribución primaria y secundaria de televisión digital, el vídeo por demanda, la videofonía, la videoconferencia, etc. Debido a la flexibilidad de las normas de codificación de vídeo, los códecs que compiten no proporcionan la misma calidad de imagen. Por ello, los métodos de evaluación de la calidad vídeo representan instrumentos importantes con los que comparar la calidad vídeo de códecs rivales y cuantificar su calidad de funcionamiento en un gran número de aplicaciones.

El reto a la hora de desarrollar técnicas de estimación de la calidad de los sistemas de compresión de vídeo procede, en parte, de que los algoritmos de compresión introducen degradaciones de vídeo que dependen mucho de los niveles de detalle y movimiento de las escenas. Además, la percepción visual de las degradaciones de vídeo depende también de los propios detalles y movimientos de las escenas. Por ello, los métodos tradicionales de evaluación, basados en señales de prueba estáticas, no son adecuados para cuantificar la calidad de funcionamiento de los sistemas de compresión de vídeo.

Este apéndice presenta una metodología de evaluación de la calidad vídeo, en la que la señal de vídeo se procesa mediante sistemas de transmisión unidireccional que utilizan interfaces digitales y, en teoría, medios de transporte digitales. El método ha sido aplicado para evaluar sistemas de compresión de vídeo de acuerdo con la norma MPEG [1] y [2], pero también se podría utilizar para evaluar otros tipos de sistemas, tales como los códecs de vídeo basados en técnicas de análisis diferentes (a saber, ondas pequeñas y filtros de predicción) y codificadores/decodificadores de señales compuestas.

La figura II.1 muestra la configuración del proceso de cálculo de los parámetros objetivos aplicado para la estimación de la calidad vídeo. El formato de fichero de las señales de vídeo digital de entrada y salida es YCbCr4:2:2, según especifica UIT-R BT.601-5 [3].

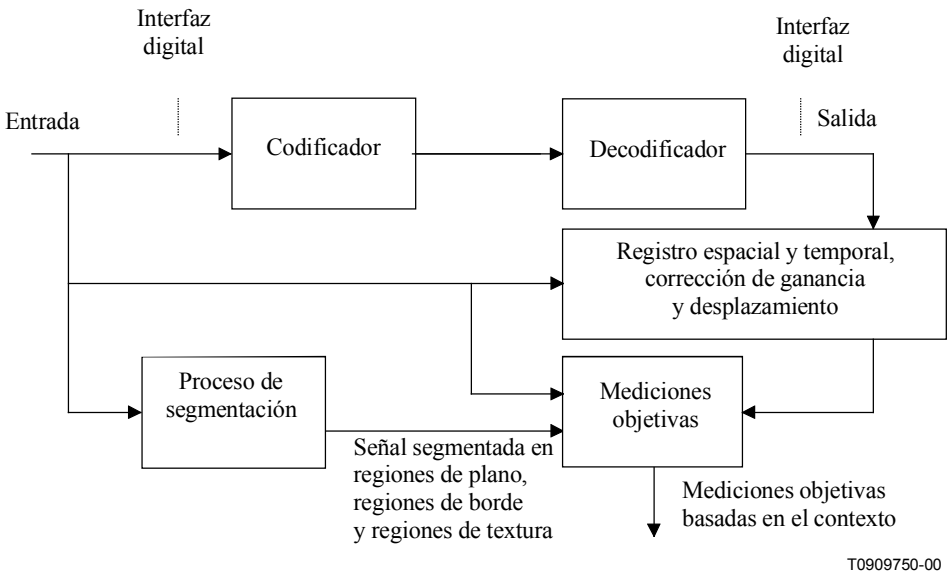


Figura II.1/J.144 – Cálculo de parámetros objetivos

En la figura II.1, cada parámetro objetivo se calcula por separado dentro de los siguientes contextos de las escenas: regiones de plano, regiones de borde y regiones de textura. Se trata de uno de los aspectos más importantes de esta metodología. La distorsión imputable a la división en bloques, por ejemplo, se puede medir mediante un detector de bordes aplicado a las regiones de plano de la escena de vídeo, en donde destaca más la percepción visual de esa distorsión. La complejidad de cálculo del método se reduce utilizando estimadores sencillos y limitando su cálculo a los contextos

correspondientes de las escenas. Dichos contextos se definen mediante un algoritmo de segmentación de la imagen que se aplica a las escenas naturales originales (es decir, a la señal de prueba de entrada). Este tipo de algoritmo comporta normalmente una complejidad de cálculo elevada, aunque sólo se ejecuta una vez. Se señala que el registro espacial y temporal entre las señales de vídeo de entrada y salida y la corrección de ganancia y desplazamiento también se requieren para calcular correctamente los parámetros objetivos. La información sobre registro (o alineación) y calibración se trata en [4].

Los parámetros objetivos se calculan por comparación directa entre escenas originales y degradadas. Todos los estimadores se aplican a campos en vez de a tramas de vídeo para asegurar la fiabilidad estadística de las mediciones en escenas de gran movimiento.

Se define un modelo basado en la percepción que predice índices subjetivos calculando la relación entre mediciones objetivas y resultados de las pruebas de evaluación subjetiva aplicadas a un conjunto de escenas naturales y códecs de vídeo MPEG-2. Estos modelos de percepción dependientes de la escena se definen en dos pasos, a saber:

1) aproximación de la relación entre cada parámetro y el nivel de degradación subjetivo mediante una curva logística, con la que se obtiene un nivel de degradación estimado para cada parámetro;

2) obtención del resultado final combinando linealmente los niveles de degradación estimados, con ponderación de cada nivel de degradación proporcional a su fiabilidad estadística.

Los detalles de la metodología que acaba de esbozarse se presentan en las cláusulas siguientes. La cláusula II.2 da una breve descripción de la configuración adoptada para las pruebas de evaluación subjetiva. Los métodos para determinar los parámetros objetivos y para segmentar las escenas naturales se describen en la cláusula II.3. La cláusula II.4 introduce los modelos perceptivos de estimación de la calidad subjetiva e informa sobre los resultados que se obtuvieron en el estudio. En esta cláusula se indican además las ventajas de utilizar parámetros objetivos basados en la segmentación de imágenes para la estimación de la calidad subjetiva y la dependencia de los modelos perceptivos con respecto a la categoría de los evaluadores y a la distancia de observación desde el monitor (4H o 6H, 4 ó 6 veces la altura de la imagen). Las conclusiones de esta contribución se presentan en la cláusula II.5.

II.2 Pruebas de evaluación subjetiva

El laboratorio de procesamiento de imágenes de CPqD/TELEBRÁS (Instituto Brasileño de Investigación y Desarrollo de las Telecomunicaciones) tiene una sala especial para realizar pruebas de evaluación subjetiva, de acuerdo con UIT-R BT.500-7 [5]. Esa sala se utilizó para evaluar la calidad de funcionamiento de algunos códecs de vídeo MPEG-2 fabricados y simulados con un subconjunto de escenas naturales propuestas por UIT-R BT.802-1 [6]. Los códecs MPEG-2 fabricados los proporcionó la compañía de televisión brasileña TV Globo. Las escenas fueron procesadas además por el programa informático de codificación MPEG-2 del que se disponía en CPqD/TELEBRÁS.

A continuación se da una breve descripción de la configuración utilizada para la evaluación subjetiva de los códecs de vídeos MPEG-2 arriba mencionados.

II.2.1 Sesiones de evaluación subjetiva

El cuadro II.1 presenta un resumen de las condiciones en que se desarrollaron las pruebas de evaluación subjetiva.

Cuadro II.1/J.144 – Condiciones de las pruebas de evaluación subjetiva

Condiciones para la evaluación	De acuerdo con 2.1 de UIT-R BT.500-7 [5]
Fuente de las señales	D1 VTR
Monitor	Monitor de estudio de 20" con interfaz digital
Distancias de observación	4H y 6H
Método de evaluación	Método escala de degradación con doble estímulo, (DSIS, *double-stimulus impairment scale*) con nueve puntos en el intervalo de 1 a 5 [5]
Secuencias de prueba	5 escenas de televisión digital con definición convencional (véase II.2.2)
Duración de la presentación	10 segundos (señal original) + 3 segundos (señal gris) + 10 segundos (señal que se evalúa) + 5 segundos para la votación, según propone la figura 3.a de UIT-R BT.500-7 [5]
Evaluadores	14 expertos y 34 no expertos
Evaluadores por sesión	5
Sesiones por evaluador	2
Presentaciones por sesión	48
Ítems evaluados	Véase II.2.3
Presentación de los resultados	Desviación media y desviación típica del nivel de degradación con respecto a la señal de referencia (escena original) Descarte de notas y evaluadores según indica UIT-R BT.500-7 [5]

II.2.2 Escenas naturales

En las sesiones de evaluación subjetiva se utilizó un conjunto de cinco escenas naturales (véase el cuadro II.2), definidas como secuencias de prueba para televisión convencional en UIT-R BT.802-1 [6]:

Cuadro II.2/J.144 – Escenas naturales utilizadas para la evaluación subjetiva

Nombre de la secuencia	Número de escena en UIT-R BT.802-1
Flower Garden	15
Mobile and Calendar	30
Table Tennis	29
Diva with Noise	17
Kiel Harbour-4	26

II.2.3 Sistemas sometidos a prueba

En la sesión de evaluación subjetiva se incluyeron, en total, 26 sistemas. Se presentan en el cuadro II.3.

Cuadro II.3/J.144 – Sistemas sometidos a prueba

Grupo	Tipo	Características	Ítems evaluados
1	Códec MP@ML MPEG-2 fabricado para aplicaciones CBR (velocidad binaria constante)	Velocidades binarias: 5, 10 y 15 Mbit/s N = 12 y M = 2	6
2	Códec MP@ML MPEG-2 simulado para aplicaciones CBR	Velocidades binarias: 2.5, 5, 7.5, 10, 12 y 15 Mbit/s N = 12 y M = 1 y 2	12
3	Códec 422P@MPL MPEG-2 fabricado para aplicaciones CBR	Velocidades binarias: 18 Mbit/s N = 2 y M = 2	1
4	Códec MP@ML MPEG-2 simulado para aplicaciones VBR (velocidad binaria variable) utilizando solamente codificación dentro de la trama	Escala de cuantificador fija [2] en 4, 8, 16, 32 y 62	5
5	Conversión de señales compuestas	NTSC y PAL-M	2

II.3 Mediciones objetivas basadas en el contexto

En esta cláusula se describe el material de vídeo utilizado para la evaluación objetiva (es decir, el material utilizado para el cálculo de los parámetros objetivos – véase II.3.1), se proponen tres métodos de segmentación de la imagen que se pueden utilizar para dividir el material de vídeo en regiones de plano, de borde y de textura (véase II.3.2) y se presentan los parámetros objetivos que han sido adoptados en este estudio (véase II.3.3).

II.3.1 Material de vídeo utilizado para la evaluación objetiva

El material de vídeo utilizado para la evaluación objetiva consiste en una secuencia de vídeo larga, de 17 segundos, compuesta por 10 cuñas de escenas naturales y 2 cuñas de señales de prueba artificiales.

De entre las escenas naturales presentadas en II.2.2 se seleccionaron cinco cuñas, de dos segundos de duración cada una. El motivo de utilizar cuñas de 2 segundos en vez de cuñas de 10 segundos, como en las pruebas subjetivas descritas en la cláusula II.2, era reducir la complejidad de cálculo del proceso de evaluación objetiva. La elección de cuñas de 2 segundos se basó en los criterios siguientes:

- La cuña de 2 segundos de una escena determinada representa un segmento crítico de su material de 10 segundos en comparación con la criticidad media de la escena. La criticidad se definió como el número de bits por trama resultante del proceso de codificación de un códec MP@ML MPEG-2 (N = 12 y M = 2) con velocidad binaria variable y escala del cuantificador igual a 16.
- La cuña de 2 segundos de esta escena representa también un segmento crítico desde el punto de vista subjetivo, cuando la escena es procesada por un códec MP@ML MPEG-2 (N = 12 y M = 2) a 5 Mbit/s.

Las otras cinco cuñas, de 1 segundo de duración cada una, consisten en escenas de bajo o nulo movimiento. Estas escenas han sido utilizadas en el proceso de evaluación objetiva, intercalando las cuñas previas de 2 segundos de duración, para probar el comportamiento adaptable de los códecs de vídeo MPEG-2 (es decir, el comportamiento con respecto al control de la velocidad y la calidad, la calidad de funcionamiento en régimen estable y en transición de escena). Se especifican también en UIT-R BT.802-1 [6]. Aunque quede fuera del alcance de la presente contribución, es importante señalar que la determinación de la variación de la calidad de funcionamiento (dispersión de la relación señal/ruido) después de cada transición de escena y en régimen estable (diferencia de

comportamiento en las tramas I, P y B) ha sido utilizada para caracterizar el comportamiento dinámico de los códecs de vídeo MPEG-2 fabricados.

Las señales de prueba artificiales son (1 segundo de duración cada una):

- Ruido de banda estrecha [4] – señal de vídeo estática y tricromática definida por ruido con una resolución de aproximadamente 1/25 del límite Nyquist e histograma aproximadamente uniforme para cada uno de los componentes Y, Cb y Cr.
- Zona circular [4] – señal de vídeo estática y tricromática definida por un diagrama sinusoidal para los componentes Y, Cb y Cr, con frecuencias horizontal y vertical constantes a lo largo de la misma columna y de la misma línea de un campo de vídeo determinado, respectivamente, y frecuencias crecientes hacia fuera a partir del centro de la imagen.

Estas señales artificiales han sido utilizadas para determinar los parámetros siguientes:

- desplazamiento de vídeo activo;
- zona de vídeo activo;
- ganancia y desplazamiento;
- respuesta en frecuencia bidimensional;
- desplazamiento entre crominancia y luminancia (se ha observado muy a menudo un desplazamiento vertical entre esos componentes en sistemas MP@ML MPEG-2 fabricados, debido a las conversiones YCbCr4:2:0 ⇔ YCbCr4:2:2, que crea un halo de chromaticidad espúrea en los bordes de la señal de salida).

En el cuadro II.4 se describe el material de prueba de 17 segundos correspondiente.

Cuadro II.4/J.144 – Material de prueba para la evaluación objetiva

Código de tiempo (mm:ss:ff)	Escena	Nombre abreviado	Característica temporal	Duración (segundos)
00:00:00	Narrow-Band Noise	Noise	estático	1
00:01:00	Flower Garden	Garden	dinámico	2
00:03:00	Tree	Tree	estático	1
00:04:00	Mobile and Calendar	Mobile	dinámico	2
00:06:00	Clown	Clown	estático	1
00:07:00	Table Tennis	Tennis	dinámico	2
00:09:00	Balls of Wool	Balls	dinámico	1
00:10:00	Diva with Noise	Diva	dinámico	2
00:12:00	Boy with Toys	Boy	estático	1
00:13:00	Kiel Harbour-4	Kiel	dinámico	2
00:15:00	Young Couple	Couple	estático	1
00:16:00	Circular Zone-Plate	Zone Plate	estático	1

II.3.2 Segmentación espacial

Se han elaborado tres algoritmos de segmentación de la imagen [7]. El primero es un algoritmo de segmentación de imagen basado en la detección de bordes utilizando filtrado recursivo (véase II.3.2.1), el segundo es un algoritmo de segmentación de imagen borrosa basado en las características espaciales (véase II.3.2.2) y el tercero es un algoritmo de segmentación de imagen basado en lo que se denomina vertiente o divisoria de aguas (véase II.3.2.3). Los resultados de la evaluación objetiva aplicando estos algoritmos se analizan en II.4.3. La estrategia de los algoritmos

de segmentación consiste en clasificar el componente luminancia de cada campo de vídeo en tres contextos mutuamente excluyentes: regiones de plano, regiones de borde y regiones de textura. A continuación se describen brevemente los algoritmos:

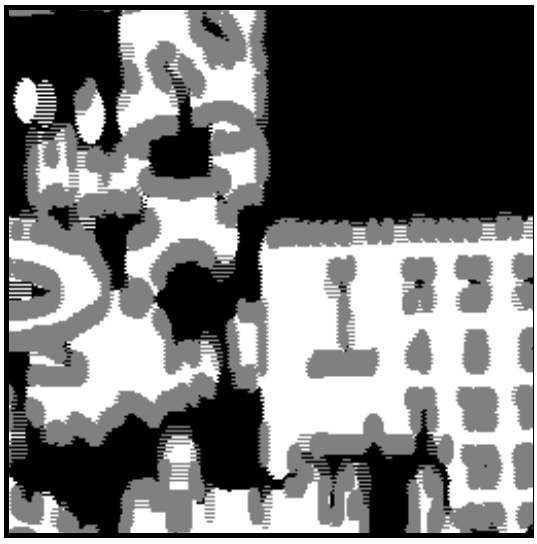
II.3.2.1 Algoritmo I: Segmentación de la imagen basada en la detección de bordes utilizando filtrado recursivo

Este algoritmo clasifica inicialmente cada píxel, en base a la variación del brillo calculada en la zona en torno al píxel, como perteneciente o no a las regiones de plano de la imagen. La imagen binaria resultante se uniformiza a continuación mediante un filtro intermedio [7]. El algoritmo aplica también a la imagen original un detector de bordes basado en el filtrado recursivo. Los bordes de la frontera de las regiones de plano se clasifican como pertenecientes a las regiones de borde. Las regiones de textura son las regiones restantes de la imagen.

La figura II.2 muestra, a título de ejemplo, parte de la escena Mobile y Calendar. El resultado de la segmentación mediante el algoritmo I de esta parte se puede ver en la figura II.3. Se señala que las regiones de plano están representadas por píxels blancos, las regiones de borde por píxels grises y las regiones de textura por píxels negros.

T0909760-00

Figura II.2/J.144 – Parte de Mobile y Calendar



T0909770-00

Figura II.3/J.144 – Resultado de la segmentación

II.3.2.2 Algoritmo II: Segmentación de la imagen borrosa basada en las características espaciales

Este algoritmo se desarrolla en dos pasos. En el primero, el algoritmo asigna una función de pertenencia, definida en el intervalo [0, 1], a cada uno de los tres contextos que se clasifican. En la función de pertenencia de las regiones de plano, el valor de pertenencia de un píxel se define como inversamente proporcional a la variación del brillo calculada dentro de la zona en torno al píxel. El gradiente morfológico [8] aplicado a esta función define la función de pertenencia de las regiones de borde. El complemento de la unión confusa [9] entre esas dos funciones de pertenencia define la función de pertenencia de las regiones de textura. En el segundo paso, cada píxel se clasifica como perteneciente al contexto con el valor más alto de entre sus tres valores de pertenencia determinados en la etapa previa.

II.3.2.3 Algoritmo III: Algoritmo de segmentación de la imagen basado en la divisoria de aguas

Este algoritmo simplifica primero el componente luminancia incrementando sus regiones homogéneas mediante la aplicación de un filtro uniformizador de preservación del borde [10]. En segundo lugar, se aplica un algoritmo de divisoria de aguas al gradiente morfológico de la imagen simplificada. La divisoria de aguas detecta regiones homogéneas, a las que se denomina cuencas colectoras, con el contraste relativo mínimo especificado. Las regiones de plano son cuencas colectoras cuya superficie es superior a un determinado umbral. Las regiones de textura vienen dadas por la erosión del complemento de las regiones de plano. Las regiones de borde son las regiones restantes de este proceso.

II.3.3 Parámetros objetivos

Los parámetros objetivos se obtienen para cada contexto (plano, borde y textura) y a partir de muestras de luminancia y crominancia de las señales de entrada (Y_{ref}, Cb_{ref} y Cr_{ref}) y salida (Y_{dec}, Cb_{dec} y Cr_{dec}), tras el registro espacial temporal y la corrección de ganancia y desplazamiento, como se muestra en la figura II.1. Las medidas y el proceso subyacente de cálculo de las mismas se describen como sigue:

- Error cuadrático medio (MSE, *mean square error*).
- Diferencia Sobel positiva (PSD, *positive Sobel difference*).
- Diferencia Sobel negativa (NSD, *negative Sobel difference*).
- Diferencia Sobel absoluta (ASD, *absolute Sobel difference*).

Sean X(i,j) el j-ésimo píxel de la i-ésima línea de la señal de entrada, Z(i,j) el j-ésimo píxel de la i-ésima línea de la señal de salida y los elementos $X_m(i,j)$ y $Z_m(i,j)$, los píxels de las señales de entrada y salida, respectivamente, después de un filtrado intermedio.

El valor del MSE en un contexto R (plano, borde o textura) se calcula aplicando la siguiente fórmula:

$$SE(i,j) = [\ X(i,j) - Z(i,j)\ ]^2, \text{ donde } (i,j) \in R.$$

El valor de la PSD en un contexto R (plano, borde o textura) se calcula aplicando la siguiente fórmula:

$$PS(i,j) = \text{máx}\ [\ \text{sobel}(X_m(i,j)) - \text{sobel}(Z_m(i,j))\ ,\ 0\ ], \text{ donde } (i,j) \in R.$$

El valor de la NSD en un contexto R (plano, borde o textura) se calcula aplicando la siguiente fórmula:

$$NS(i,j) = -\ \text{máx}\ [\ \text{sobel}(Z_m(i,j)) - \text{sobel}(X_m(i,j))\ ,\ 0\ ], \text{ donde } (i,j) \in R.$$

El valor de la ASD en un contexto R (plano, borde o textura) se calcula aplicando la siguiente fórmula:

$$AS(i,j) = |\ \text{sobel}(X_m(i,j)) - \text{sobel}(Z_m(i,j))\ |, \text{ donde } (i,j) \in R.$$

En otras palabras, ASD = PSD + NSD.

Los parámetros objetivos utilizados para la estimación de la calidad subjetiva se refieren al valor medio de las medidas antes mencionadas calculadas con un conjunto de 2N campos del tramo final (esto es, en condición de régimen estable) de cada una de las cinco cuñas indicadas en el cuadro II.4 y que pertenecen a las escenas presentadas para la evaluación subjetiva. El valor N es un múltiplo del intervalo entre imágenes codificadas dentro de trama (tipo I), es decir, es un múltiplo de la longitud del grupo de imágenes (GOP, *group of picture*) [1, 2]. Para que esta condición la satisfagan todos los sistemas definidos en el cuadro II.3, se elige N = 12.

II.4 Estimación de la calidad subjetiva

En esta cláusula se describe la manera de definir los modelos de estimación de la calidad subjetiva para cada escena. La cláusula II.4.1 describe un modelo perceptivo para estimar el nivel de degradación subjetivo en base a un solo parámetro. Los resultados de esta aproximación para cada parámetro subjetivo se combinan linealmente para estimar el nivel de degradación subjetivo final. Este modelo de predicción lineal se presenta en II.4.2. En la cláusula II.4.3 se presentan y analizan los resultados de este estudio.

II.4.1 Estimación de la calidad subjetiva en base a un solo parámetro: Aproximación logística

Para cada escena, la relación entre cada parámetro objetivo D y el resultado subjetivo U se define inicialmente según se indica a continuación:

El nivel de degradación normalizado entre 0% y 100% viene dado por [5]:

$$d = (U_{\text{máx}} - U)/(U_{\text{máx}} - U_{\text{mín}}) \times 100\%$$

La relación entre d y cada parámetro objetivo D se aproxima mediante la función logística siguiente [5]:

$$\underline{d} = \frac{1}{1 + \left(\frac{D_M}{D}\right)^G} \times 100\%$$

donde los valores D_M y G se calculan de manera que se reduzca al mínimo el error cuadrático medio:

$$e = E\left[\{d - \underline{d}\}^2\right]$$

para cada escena y cada parámetro objetivo separadamente. La fiabilidad estadística de $\underline{d}$ se define como 1/e.

II.4.2 Estimación de la calidad subjetiva: Predicción lineal en tres etapas

La estimación del nivel de degradación normalizado d mediante un conjunto de niveles de degradación estimados $\underline{d}$ (uno por parámetro, como se define en II.4.1) se implementa en tres etapas de predicción lineal, descritas a continuación.

Etapa 1

En primer lugar se consideran los conjuntos de niveles de degradación estimados siguientes, seleccionados para el componente luminancia:

- $\underline{d}^{MSE}$
- $\underline{d}^{PSD}$ y $\underline{d}^{NSD}$
- $\underline{d}^{ASD}$
- $\underline{d}^{MSE}$, $\underline{d}^{PSD}$ y $\underline{d}^{NSD}$
- $\underline{d}^{MSE}$ y $\underline{d}^{ASD}$

Para una escena dada y un contexto de esa escena (plano, borde o textura), el mejor conjunto es aquel cuyo error de predicción sea menor. Utilizando ese criterio para elegir un conjunto de niveles de degradación estimados para cada contexto, esta etapa combina linealmente los niveles de degradación de cada conjunto seleccionado y genera tres valores de estimación (uno por contexto) indicados por: $\underline{d}_{YP}$, $\underline{d}_{YE}$ y $\underline{d}_{YT}$.

De manera similar, los conjuntos considerados de niveles de degradación estimados para los componentes crominancia de la escena son:

- $\underline{d}^{MSE(Cb)}$ y $\underline{d}^{MSE(Cr)}$
- $\underline{d}^{ASD(Cb)}$ y $\underline{d}^{ASD(Cr)}$

y los valores de la estimación resultantes (uno por contexto) se indican por: $\underline{d}_{CP}$, $\underline{d}_{CE}$ y $\underline{d}_{CT}$.

Etapa 2

Los valores de estimación $\underline{d}_P$, $\underline{d}_E$ y $\underline{d}_T$ se obtienen a partir de la predicción lineal basada en los vectores ($\underline{d}_{YP}$, $\underline{d}_{CP}$), ($\underline{d}_{YE}$, $\underline{d}_{CE}$) y ($\underline{d}_{YT}$, $\underline{d}_{CT}$), respectivamente.

Etapa 3

Los valores de estimación $\underline{d}_P$, $\underline{d}_E$ y $\underline{d}_T$ se combinan aplicando una predicción lineal para generar el nivel de degradación estimado $\underline{d}$.

En todas las etapas anteriores, los predictores se atienen a las restricciones que se indican a continuación.

Sea ($\underline{d}_1$, $\underline{d}_2$, ..., $\underline{d}_P$) el vector de entrada del predictor lineal. La salida $\underline{d}_o$ viene dada por:

$$\underline{d}_o = \sum a_i \underline{d}_i$$

calculándose las ponderaciones $\{a_i\}$ de modo que se reduzca al mínimo el error cuadrático medio:

$$E\left[\{d - \underline{d}_o\}^2\right], \text{ con lo que}$$

$$\sum a_i = 1 \text{ y}$$

$$a_i / a_k = e_k / e_i$$

donde la fiabilidad estadística de $\underline{d}_i$ es $1/e_i$, como se define en II.4.1.

Se ha observado que este tipo de predicción es más consistente que la que se obtiene mediante predictores óptimos, porque depende menos de la base de datos de entrenamiento. Se consiguen mejores resultados cuando se aplica a bases de datos de prueba, de lo cual se da un ejemplo en II.4.3.

II.4.3 Estimación de la calidad subjetiva: Presentación y análisis de los resultados

Esta cláusula se refiere a tres temas principales. Los resultados y los modelos de predicción obtenidos mediante la estimación de la calidad subjetiva basada en el algoritmo I (el algoritmo de segmentación de la imagen descrito anteriormente en II.3.2.1) se presentan en II.4.3.1. En la cláusula II.4.3.1 se expone además la dependencia entre los modelos perceptivos y la categoría de los evaluadores (expertos y no expertos) y entre los modelos perceptivos y la distancia de observación desde el monitor (4H y 6H). La variación de la exactitud de la estimación con los algoritmos de segmentación de la imagen se analiza en II.4.3.2. En la cláusula II.4.3.3 se indican las ventajas del método de estimación subjetiva propuesto en comparación con otros métodos basados en mediciones globales o de predicción óptima.

II.4.3.1 Resultados: Modelos perceptivos y calidad de funcionamiento

El cuadro II.5 presenta los resultados del método de estimación subjetiva basado en el algoritmo I (véase II.3.2.1) con la segmentación de las escenas siguientes: Garden, Mobile, Tennis, Diva y Kiel, separadamente. En el cuadro II.5:

- Las ponderaciones de la predicción lineal a la que se refiere la etapa 2 de II.4.2 son equivalentes a las ponderaciones subjetivas relativas de las degradaciones de la luminancia (Y) y la crominancia (C) en las regiones de plano, regiones de borde y regiones de textura. El valor medio global calculado con todas las escenas se indica en la última línea del cuadro.
- Las ponderaciones de la predicción lineal a la que se refiere la etapa 3 de II.4.2 son equivalentes a las ponderaciones subjetivas relativas de la degradación en las regiones de plano (P), regiones de borde (E) y regiones de textura (T). El valor medio global calculado con todas las escenas se indica en la última línea del cuadro.
- El error cuadrático medio (MSE) y el error medio absoluto (MAE, *mean absolute error*) entre el nivel de degradación normalizado d y el nivel de degradación estimado $\underline{d}$, teniendo en cuenta una escala de normalización de 0% a 100%, se muestran en las dos últimas columnas del cuadro. El error entre el nivel de degradación normalizado medio y el nivel de degradación estimado medio, calculados con todas las escenas, se muestra en la última línea de estas columnas.

Los resultados presentados en el cuadro II.5 se refieren a los modelos perceptivos obtenidos a partir de las notas subjetivas de los 34 evaluadores no expertos del cuadro II.1 y de los 26 sistemas evaluados del cuadro II.3.

Cuadro II.5/J.144 – Modelos perceptivos y resultados: Evaluadores no expertos

	Etapa 2: Plano		**Etapa 2: Borde**		**Etapa 2: Textura**		**Etapa 3**			**Error**	
Escena	**Y(%)**	**C(%)**	**Y(%)**	**C(%)**	**Y(%)**	**C(%)**	**P(%)**	**E(%)**	**T(%)**	**MSE**	**MAE**
Garden	61	39	70	30	37	63	13	37	51	18,1	3,0
Mobile	74	26	75	25	63	37	83	7	9	24,2	3,6
Tennis	67	33	65	35	70	30	45	13	42	25,3	3,5
Diva	49	51	92	8	42	58	27	59	14	5,4	1,5
Kiel	62	38	66	34	40	60	32	39	29	22,7	3,6
Global	**63**	**37**	**73**	**27**	**50**	**50**	**40**	**31**	**29**	**6,2**	**1,8**

Los cuadros II.6 y II.7 muestran la dependencia entre modelos perceptivos y resultados para:

- evaluadores no expertos y evaluadores expertos;
- distancia de observación (4H y 6H) desde el monitor (en cada caso con el 50% del número total de evaluadores).

Cuadro II.6/J.144 – Modelos perceptivos y resultados: Evaluadores no expertos y evaluadores expertos

	Evaluadores no expertos						Evaluadores expertos					
	Región			Componente		Error	Región			Componente		Error
Escena	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**
Garden	13	37	51	52	48	18,1	12	53	34	51	49	23,5
Mobile	83	7	9	73	27	24,2	72	13	15	72	28	73,4
Tennis	45	13	42	68	32	25,3	47	12	41	70	30	48,1
Diva	27	59	14	73	27	5,4	22	42	36	55	45	21,2
Kiel	32	39	29	57	43	22,7	43	36	21	47	53	44,1
Global	**40**	**31**	**29**	**62**	**38**	**6,2**	**39**	**31**	**30**	**58**	**42**	**12,1**

Cuadro II.7/J.144 – Modelos perceptivos y resultados: Distancias de observación de 6H y 4H

	Distancia de observación de 6H						Distancia de observación de 4H					
	Región			Componente		Error	Región			Componente		Error
Escena	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**
Garden	15	40	45	51	49	20,0	9	35	55	47	53	16,8
Mobile	83	8	9	77	23	24,7	71	13	16	62	38	59,4
Tennis	47	15	38	60	40	40,9	42	10	48	74	26	21,4
Diva	41	36	22	64	36	15,5	21	46	33	54	46	12,5
Kiel	34	40	26	54	46	18,3	31	46	23	57	43	26,0
Global	**44**	**28**	**28**	**61**	**39**	**7,9**	**35**	**30**	**35**	**59**	**41**	**9,6**

A continuación se comentan los resultados presentados en los cuadros II.5, II.6 y II.7:

- La estimación de calidad subjetiva utilizando parámetros objetivos basados en la segmentación de imagen, calculada con los 26 sistemas descritos en II.2.3, dio como resultado un error medio absoluto (MAE) de menos del 4% para cada una de las escenas y un MAE global del 1,8%, interviniendo evaluadores no expertos.
- Comparando los modelos perceptivos basados en la opinión de evaluadores expertos y no expertos, la importancia de las degradaciones de la crominancia resulta ser ligeramente superior en los modelos basados en evaluadores expertos.
- Comparando los modelos perceptivos basados en distancias de observación de 4H y 6H, la importancia de las degradaciones en las regiones de borde y de textura resulta ser notablemente superior en los modelos basados en distancias de observación de 4H, como era de prever.

II.4.3.2 Variación de la exactitud de la estimación con el algoritmo de segmentación de la imagen

En el cuadro II.8 se muestran los resultados de la estimación de la calidad subjetiva, basados en los algoritmos II y III (descritos brevemente en II.3.2), y obtenidos a partir de las notas de los 34 evaluadores no expertos. Comparando los resultados del cuadro II.8 con los resultados presentados previamente en el lado izquierdo del cuadro II.6 (para el algoritmo I), la exactitud de la estimación muestra pequeñas variaciones dependientes del algoritmo de segmentación de la imagen para una escena dada. Por otro lado, no hay ninguna variación importante en la exactitud de la estimación

global cuando se consideran los tres algoritmos de segmentación de la imagen. Cabe pensar, por todo ello, que incluso con algoritmos de segmentación de la imagen más sencillos se pueden obtener resultados satisfactorios.

Cuadro II.8/J.144 – Modelos perceptivos y resultados: Algoritmos II y III

	Algoritmo II						Algoritmo III					
	Región			Componente		Error	Región			Componente		Error
Escena	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**
Garden	9	32	59	53	47	18,0	10	44	46	53	47	15,8
Mobile	65	26	9	59	41	20,7	82	11	6	60	40	18,7
Tennis	54	27	19	70	30	28,5	68	21	11	72	28	31,3
Diva	25	50	24	75	25	7,1	50	31	19	60	40	7,4
Kiel	23	31	46	64	36	25,9	28	33	38	59	41	22,4
Global	**35**	**33**	**31**	**66**	**34**	**7,4**	**48**	**28**	**24**	**63**	**37**	**6,5**

II.4.3.3 Ventajas del método de estimación de la calidad subjetiva adoptado

El ejemplo ilustrado en el cuadro II.9 destaca dos propiedades muy importantes de cualquier método de estimación de la calidad subjetiva basado en parámetros objetivos: su exactitud y su solidez [11] a [14]. En este ejemplo se compara el método de estimación de la calidad subjetiva adoptado, que depende de las mediciones objetivas basadas en el contexto, y el método de predicción lineal descrito en II.4.2, con los métodos siguientes:

- un método que depende de las mismas mediciones objetivas basadas en el contexto, pero que utiliza la predicción óptima;
- un método que utiliza la predicción lineal descrita en II.4.2, pero adopta mediciones globales.

En el ejemplo se utilizaron los sistemas del grupo 2 y el sistema NTSC del grupo 5 como base de datos de entrenamiento, y los sistemas del grupo 1 y el sistema PAL-M del grupo 5 como base de datos de prueba (véase el cuadro II.3). Los parámetros objetivos utilizados en este ejemplo fueron los basados en el MSE y descritos en II.3.3. Los resultados se obtuvieron a partir de las notas de los evaluadores no expertos. Los valores de entrada del cuadro son errores de predicción cuadráticos medios. En la última línea se muestra el valor medio de este parámetro calculado con el conjunto de escenas.

Cuadro II.9/J.144 – Comparación: Solidez y exactitud

	Método adoptado		Predictor óptimo		Mediciones globales	
Escena	**Entrenamiento**	**Prueba**	**Entrenamiento**	**Prueba**	**Entrenamiento**	**Prueba**
Garden	3,9	87,6	2,8	71,8	3,9	62,3
Mobile	30,1	48,6	10,5	82,1	179,1	162,5
Tennis	10,8	91,3	7,7	335,0	108,9	221,2
Diva	1,4	8,9	0,8	17,7	1,8	34,3
Kiel	22,4	9,3	20,5	13,4	30,6	27,7
Valor medio	**13,7**	**49,1**	**8,5**	**104,0**	**64,9**	**101,6**

La ventaja de calcular parámetros objetivos en base al contexto se hace evidente cuando el procedimiento descrito en las cláusulas II.4.1 y II.4.2 se aplica también a mediciones globales. Se señala que la utilización de mediciones basadas en el contexto puede mejorar notablemente los resultados de la estimación en todas las escenas (excepto la de Flower Garden). Esto indica, posiblemente, que el proceso de segmentación de la imagen para Flower Garden ha de ser perfeccionado.

El ejemplo muestra también que el proceso de predicción descrito en II.4.2 es más sólido (es decir, depende menos de la base de datos de entrenamiento) cuando se compara con el predictor óptimo, mejorando los resultados de la predicción de la base de datos de prueba.

II.5 Conclusiones

Este apéndice expone una metodología de estimación de la calidad subjetiva en la que se utilizan parámetros objetivos basados en la segmentación de la imagen. Los parámetros objetivos se calculan dentro de las regiones de plano, regiones de borde y regiones de textura resultantes del proceso de segmentación de la imagen.

Los resultados presentados en este apéndice muestran que la utilización de parámetros objetivos basados en el contexto comparados con los parámetros globales da lugar a predicciones más exactas. Este aspecto viene confirmado por la utilización del modelo perceptivo basado en el método de predicción lineal que se describe en II.4.2. El método descrito lleva a resultados de predicción más consistentes cuando se compara con la predicción óptima.

Los resultados pueden mejorarse todavía si:

- se incluye la información temporal en el proceso de segmentación de la imagen (por ejemplo, las regiones de borde podrían subclasificarse aún en regiones de borde con poco movimiento y regiones de borde con mucho movimiento);
- se consideran además las regiones de plano, borde y textura de la crominancia en el proceso de segmentación de la imagen, ya que los algoritmos I, II y III se utilizaron para segmentar el componente luminancia solamente.

Por todo lo anterior, se propone la inclusión del método de predicción lineal presentado en este apéndice y las mediciones objetivas basadas en el contexto en nuevas Recomendaciones de la UIT, relacionadas con la evaluación objetiva de la calidad vídeo.

II.6 Referencias

[1] ISO/CEI 11172-1:1993, *Information technology – Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbit/s – Part 1: Systems*.

[2] UIT-T H.262 (2000), *Tecnología de la información – Codificación genérica de imágenes en movimiento e información de audio asociada: Vídeo*.

[3] UIT-R BT.601-5 (1995), *Parámetros de codificación de televisión digital para estudios con formatos de imagen normal 4:3 de pantalla ancha 16:9*.

[4] ANSI T1.801.03 (1996), *Digital transport of one-way video signals – Parameters for objective performance assessment*.

[5] Recomendación UIT-R BT.500-7 (1995), *Metodología para la evaluación subjetiva de la calidad de las imágenes de televisión*.

[6] UIT-R BT.802-1 (1994), *Imágenes y secuencias de prueba para las evaluaciones subjetivas de códecs digitales que cursan señales producidas conforme a la Recomendación UIT-R BT.601*.

[7] GONZALEZ, WINTZ (P.), Digital Image Processing, *Addison Wesley*, 1987.

[8] DOUGHERTY: An Introduction to Morphological Image Processing, *SPIE Optical Engineering Press*, Bellingham, WA, Vol. TT9, 1992.

[9] KAUFMANN (A.): Introduction to The Theory of Fuzzy Subsets, *Academic Press*, New York, NY, Vol. 1, 1975.

[10] BARRERA (J.), BANON (J.F.), LOTUFO (R.A.): Mathematical Morphology Toolbox for the Khoros System, *Conference on Image Algebra and Morphological Image Processing, V International Symposium on Optics, Imaging, and Instrumentation, SPIE's Annual Meeting*, San Diego, USA, 24-29 de julio de 1994.

[11] ITU-T Contribution COM 12-66, *Selections from the draft American National Standard – Digital transport of one-way signals – Parameters for objective performance assessment*, USA, enero de 1996.

[12] ITU-T Study Group 12 Delayed Contribution D021, *Objective and subjective measures of MPEG video quality: summary of experimental results*, USA, abril de 1997.

[13] ITU-T Study Group 12 Delayed Contribution D101, *A Two-Stage Objective Model for Video Quality Evaluation*, Bellcore, mayo de 1996.

[14] ANSI T1A1 Contribution Number T1A1.5/96-121, *Objective and subjective measures of MPEG video quality*, GTE Labs., NTIA/ITS, octubre de 1996.

APÉNDICE III

Tektronix/Sarnoff

Introducción

Los nuevos servicios de televisión digital crean una demanda de supervisión de la calidad de servicio con instrumentos de medición que son muy diferentes de sus correspondientes analógicos. Es requisito fundamental el que las mediciones de la calidad de imagen objetivas estén estrechamente correlacionadas con la evaluación de la calidad subjetiva.

En el presente apéndice se describe un instrumento de medición basado en el modelo de visión humana que puede ser utilizado dentro de un sistema de televisión digital. Su aplicación práctica da resultados que demuestran la existencia de un alto grado de correlación con las evaluaciones subjetivas efectuadas de acuerdo con UIT-R BT.500-7.

Temas tratados en concreto son:

- La medición del índice de calidad de imagen (PQR) objetivo en las redes de transmisión de vídeo digital.
- Los requisitos del preprocesamiento del vídeo antes del análisis.
- Detalles del algoritmo utilizado para efectuar el análisis.
- Los resultados de las pruebas que muestran la correlación entre mediciones objetivas y evaluaciones de la calidad de imagen subjetivas.

Se expone una solución al tema de la medición de la percepción de la calidad vídeo dentro de sistemas de vídeo digital, incorporada en una implementación comercial.

III.1 Índice de calidad de imagen (PQR) objetivo en entornos operacionales

Es bien sabido que las mediciones objetivas de la calidad de imagen se pueden efectuar con mayor exactitud cuando se dispone de información sobre el vídeo de referencia. En el diagrama genérico mostrado en la figura III.1, el vídeo de programa entra en el sistema de transmisión (vídeo de referencia) y es transmitido a través del sistema y supervisado a la salida (vídeo procesado). El análisis de las diferencias entre el vídeo procesado y el vídeo de referencia con un modelo de visión

humana proporciona una medida precisa del índice de calidad de imagen (PQR) objetivo. [En la cláusula III.4, se dan detalles del algoritmo utilizado en el modelo de visión humana Sarnoff/Tektronix desarrollado conjuntamente, con el que se obtienen índices de calidad de imagen (PQR) objetivos.]

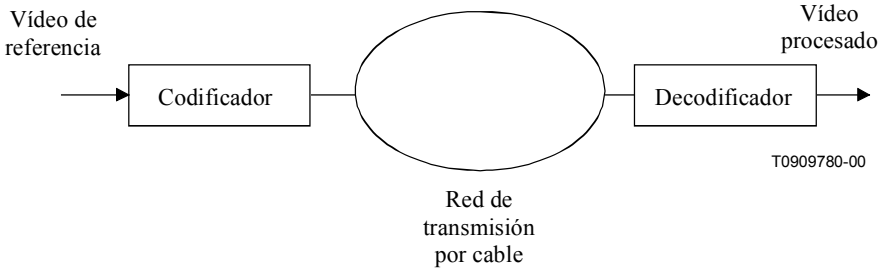


Figura III.1/J.144 – Diagrama genérico de un sistema de transmisión de vídeo comprimido

Un prerrequisito fundamental para efectuar el análisis del modelo de visión humana consiste en la normalización del vídeo procesado. La codificación y la decodificación pueden causar arrastres y desplazamientos horizontales y verticales de la imagen, y también cambios en la ganancia y el nivel de luminancia y crominancia. Se han de normalizar antes de la aplicación del modelo de visión humana. (En la cláusula III.2 se dan detalles del proceso normalizador.)

Determinados sistemas de transmisión quizás requieran que se amplíe el diagrama genérico para reflejar la existencia de códecs concatenados y/o la utilización de la codificación y decodificación PAL, pero los principios siguen siendo los mismos, y el proceso de evaluación del índice de calidad de imagen (PQR) objetivo sigue siendo válido.

En un entorno de laboratorio, se pueden utilizar secuencias de prueba de vídeo en vez de material de vídeo en directo. Representan una fuente repetible de vídeo y facilitan la realización de mediciones comunes entre diferentes laboratorios. La inclusión de una amplia gama de material de programa estándar dentro de las secuencias de vídeo de prueba garantiza la óptima validez del material de programa en directo.

En un entorno operativo, las secuencias de prueba de vídeo pueden ser sustituidas por material de programa en directo. El vídeo de referencia y el vídeo procesado procedente de un decodificador situado en la fuente de transmisión, como en la figura III.2, proporcionan una medida del índice de calidad de imagen (PQR) objetivo del sistema operacional.

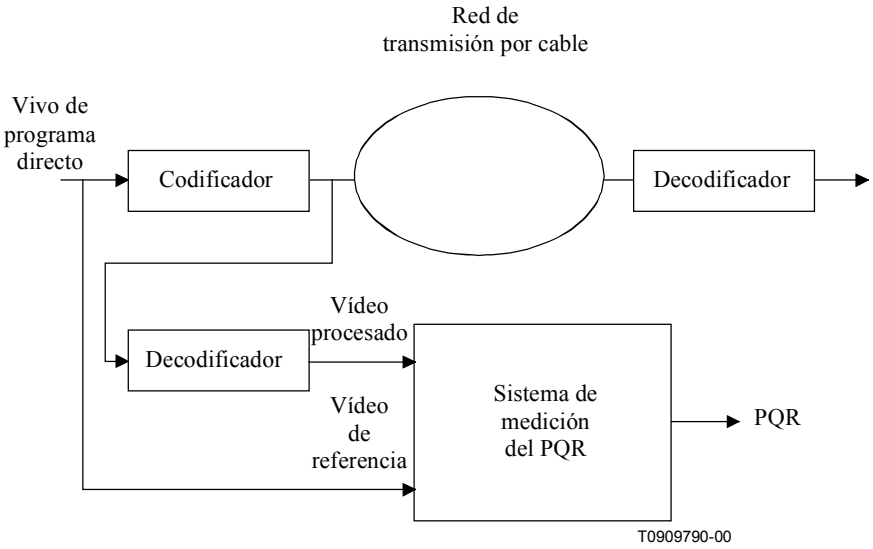


Figura III.2/J.144 – Índice de calidad de imagen (PQR) objetivo con referencia disponible

La falta de acceso al vídeo de referencia quizá restrinja la utilización de material en directo de manera continua. Un ejemplo obvio es la supervisión de la entrada de señales por satélite en un extremo de cabecera de cable. En tal situación, se puede elegir una secuencia de vídeo común, por ejemplo un logotipo de estación, como vídeo de referencia. Podría ser proporcionada por el operador de cable y almacenada localmente como referencia para la comparación con el vídeo procesado. Véase la figura III.3.

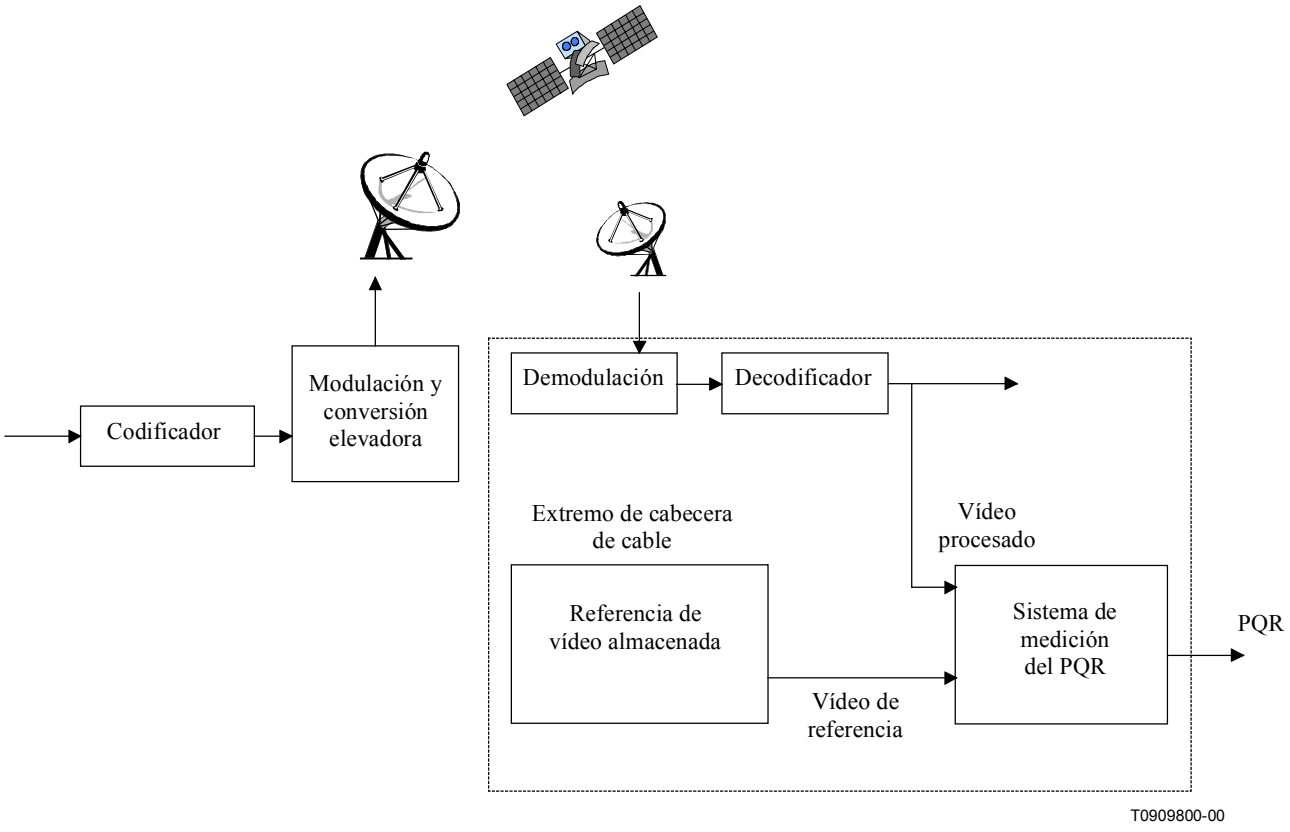


Figura III.3/J.144 – Índice de calidad de imagen (PQR) objetivo en un extremo de cabecera distante con referencia almacenada

III.2 Preprocesamiento de vídeo – Normalización

La aplicación del método índice de calidad de imagen (PQR) objetivo a cualquier sistema de vídeo requiere la normalización del vídeo procesado. Normalización significa que los cambios sistemáticos invariables con el tiempo en el vídeo desde la entrada de referencia hasta la salida del vídeo procesado se eliminan antes de efectuar la medición basada en el sistema de visión humana (HVS). En tanto que método de medición de la calidad de imagen objetiva más sensible y exacto, el método PQR se basa en filtros HVS que comparan las imágenes de referencia y las imágenes procesadas realmente píxel por píxel. Es necesario separar la medición en dos partes, normalización y cálculo del PQR, para obtener los resultados más significativos.

Los parámetros que se han de ajustar mediante el proceso de normalización son los de desplazamientos horizontales y verticales de la imagen; los cambios de ganancia de luminancia y color; los cambios de nivel de c.c. de luminancia y color; y la diferencia de retardo canal a canal de componente o luminancia a color. Puesto que estos cambios podrían producir modificaciones en la calidad de imagen percibida, deberán notificarse como parte de los resultados del método de medición. Es necesario separar esos cambios de los cálculos del PQR por dos motivos. El principal es para proporcionar el valor de PQR más exacto. El segundo motivo es que esa normalización se corresponde estrechamente con el funcionamiento de un sistema típico por lo que se refiere a los parámetros de ganancia y nivel de c.c. cuando se dispone por lo general de los ajustes apropiados y se efectúan de manera rutinaria. Normalmente se considera que pequeños desplazamientos de la imagen, horizontales o verticales, no cambian la calidad de imagen percibida, si bien su presencia es de hecho un error de la imagen y provocará problemas importantes en aplicaciones multigeneración. Los cambios de las imágenes variables con el tiempo debidos al contenido del vídeo y al sistema de compresión se miden mediante el cálculo del PQR.

La idea de normalizar antes de efectuar una evaluación de la calidad de imagen también ha de estar presente en las normas relativas a la medición subjetiva, como reflejo de lo que se entiende por funcionamiento de un sistema típico. En UIT-T P.911: Métodos de evaluación subjetiva de la calidad audiovisual para aplicaciones multimedios, y en UIT-T P.910: Métodos de evaluación subjetiva de la calidad vídeo para aplicaciones multimedios, se ha de incluir la siguiente declaración: "Los parámetros operacionales, tales como el nivel de señal, de las secuencias de prueba deberán concordar con los de las señales de alineación utilizadas para verificar las condiciones de observación [y escucha]. Cualquier ajuste operativo efectuado a fin de que la secuencia fuente o la secuencia procesada cumplan este requisito deberá ser notificado".

La figura III.4 muestra el funcionamiento del sistema del PQR con respecto a la normalización. El vídeo procesado se normaliza campo por campo por comparación con el vídeo de referencia o midiendo señales de prueba calibradas insertadas en la secuencia de referencia. Sólo se eliminan del vídeo los cambios estáticos invariables con el tiempo; los cambios dinámicos debidos a los procesos de compresión y descompresión se miden como parte del cálculo del PQR. La normalización del vídeo procesado antes de los cálculos del PQR deberá respetar los márgenes de tolerancia que se indican en el cuadro III.1.

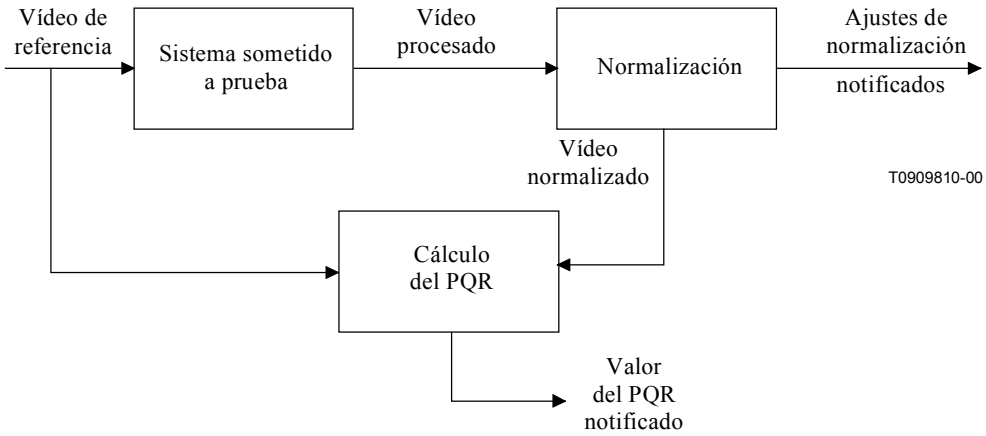


Figura III.4/J.144 – Funcionamiento del sistema de medición del PQR

Cuadro III.1/J.144 – Parámetros y tolerancia de la normalización

Parámetro	Tolerancia de la normalización
Ganancia de luminancia	< 0,2 dB
Ganancia de color (diferencia)	< 0,2 dB
Nivel de c.c. de luminancia	< 0,5% del máximo de la señal
Nivel de c.c. de color (diferencia)	< 0,5% del máximo de la señal
Diferencia de retardo canal a canal	< 2 ns
Desplazamiento de píxel horizontal	< 0,1 píxel
Desplazamiento de línea vertical	< 0,1 línea

III.3 Visión general del sistema

Los índices de calidad de imagen (PQR) objetivos son uno de los resultados clave que proporciona el sistema de medición del PQR al que se ha hecho referencia más arriba. En lo que sigue se hace una descripción del modelo de visión humana utilizado en el sistema de medición del PQR.

El modelo de visión humana Sarnoff/Tektronix es un método de predicción de los índices perceptivos que las personas asignarán a una secuencia de imágenes en color degradadas con respecto a sus correspondientes no degradadas. El modelo utiliza dos secuencias de imágenes y produce varias estimaciones de la diferencia, incluyendo una medida única de las diferencias percibidas entre las dos secuencias. Esas diferencias se cuantifican en unidades de la diferencia apenas perceptible (JND, *just-noticeable difference*) por el ser humano. Lubin (1993, 1995) ha descrito una versión del modelo que se aplica solamente a imágenes estáticas y acromáticas.

El modelo de visión humana puede ser de utilidad en un contexto general (véase la figura III.5). Una secuencia de vídeo de entrada pasa a través de dos canales diferentes hacia un observador humano (no mostrado en la figura). Uno de los canales no está distorsionado (el canal de referencia) y en el otro se distorsiona de alguna manera la imagen (el canal sometido a prueba). La distorsión, efecto lateral de alguna medida tomada por razones económicas, puede producirse en un codificador antes de la transmisión, en el propio canal de transmisión, o en el proceso de decodificación. En la figura III.5, la casilla "sistema sometido a prueba" se refiere esquemáticamente a cualquiera de esas alternativas. Por lo general, la evaluación de la calidad subjetiva de la imagen de prueba con respecto a la secuencia de referencia requerirá un observador humano y un dispositivo de visualización real. La evaluación se facilitaría sustituyendo la visualización y el observador por el modelo de visión humana, que compara las secuencias de prueba y referencia para producir una secuencia de mapas JND en lugar de la comparación subjetiva.

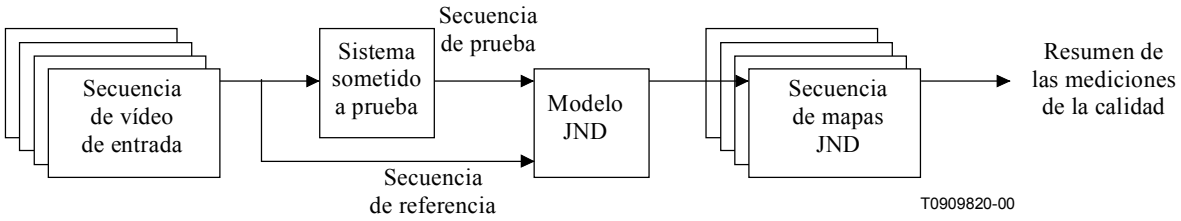


Figura III.5/J.144 – Modelo de visión humana en la evaluación de un sistema

La figura III.6 muestra una visión general del algoritmo. Las entradas son dos secuencias de imágenes de una longitud cualquiera. Por cada campo de una secuencia de entrada hay tres conjuntos de datos, rotulados Y', C_b' y C_r' en la parte superior de la figura III.6, obtenidos a partir, por ejemplo, de una cinta D1. Los datos Y, C_b y C_r son transformados a continuación en las tensiones de cañón electrónico R', G' y B' que generan los valores de píxels presentados visualmente. En el modelo, las tensiones R', G' y B' son objeto de un procesamiento ulterior para transformarlas en una señal de luminancia y dos imágenes cromáticas que pasan a las etapas subsiguientes.

La finalidad del procesamiento de extremo delantero es transformar las señales de entrada de vídeo en salidas de luz, y transformar a continuación esas salidas de luz en cantidades definidas psicofísicamente que caracterizan por separado luma y chroma.

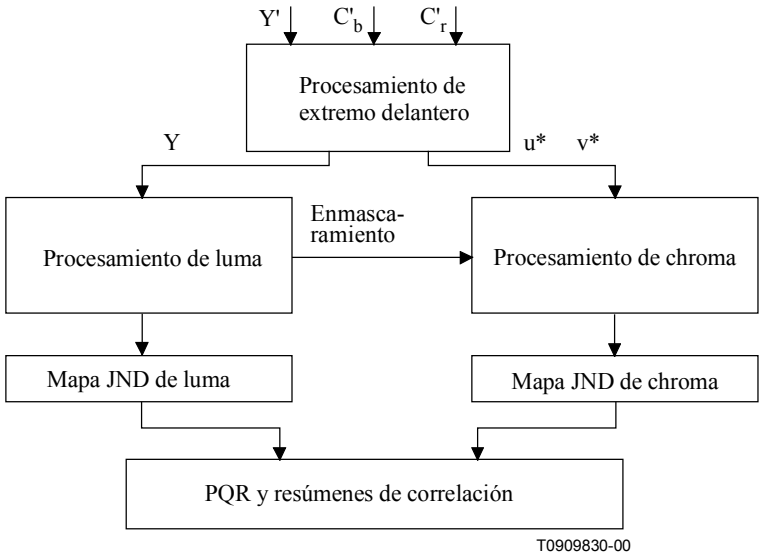


Figura III.6/J.144 – Diagrama del modelo de visión humana de Sarnoff/Tektronix

La etapa procesamiento de luma acepta dos imágenes (de prueba y de referencia) de luminancias Y, expresadas como funciones de la luminancia máxima de la visualización. A partir de esas entradas, la etapa de procesamiento de luma genera un mapa JND de luma. Ese mapa es una imagen cuyos niveles de gris son proporcionales al número de diferencias apenas perceptibles (JND) entre la imagen de prueba y la imagen de referencia en la ubicación del píxel correspondiente.

Un procesamiento similar, basado en el espacio de color uniforme CIE L*u*v*, tiene lugar por cada una de las imágenes de chroma u* y v*. Los resultados del procesamiento de u* y v* se combinan para producir un mapa JND de chroma. Tanto el procesamiento de chroma como el de luma se ven influidos por las entradas procedentes del canal de luma llamado *enmascaramiento*, que hace que las diferencias percibidas sean más o menos visibles dependiendo de la estructura de las imágenes de luma.

Cada uno de los mapas JND de luma, chroma y luma-chroma combinados está disponible, como resultado, junto con un pequeño número de medidas resumidas derivadas de esos mapas. Valores resumen de PQR únicos modelan la evaluación global por parte de un observador de las distorsiones presentes en una secuencia de prueba. Los mapas JND dan una visión más detallada de la ubicación y la importancia de los efectos visuales secundarios.

Se señala que, tras el modelo aquí presentado, subyacen dos hipótesis básicas:

a) Cada píxel tiene forma cuadrada y subtiende un ángulo de observación de 0,3º. Este valor se obtuvo con una pantalla cuya altura era de 480 píxels, y una distancia de observación de cuatro veces la altura de la pantalla (la distancia de observación más cercana prescrita por UIT-R BT.500). Cuando el modelo se compara con la percepción humana con distancias de observación superiores a la de cuatro alturas de pantalla, se comprueba que sobreestima la sensibilidad humana a los detalles espaciales. En ausencia de limitaciones impuestas a la distancia de observación, el modelo se elige de modo que sea tan sensible como se pueda dentro del marco de UIT-R BT.500.

b) El modelo es aplicable con luminancias de pantalla de 0,01 a 100 pies-lumen (para las que se calibró la sensibilidad global), pero con una precisión máxima de unos 20 pies-lumen (para la que se calibraron todas las frecuencias espacio-temporales). Se supone que una luminancia cambiante provoca cambios proporcionales de la sensibilidad a todas las frecuencias espacio-temporales, y esta suposición es menos importante cerca de 20 pies-lumen, en donde se llevó a cabo la mayor parte de la calibración.

El procesamiento que se muestra en algunas de las casillas de la figura III.6 se describe con más detalle a continuación, en las figuras III.7, III.8 y III.9.

III.4 Visión general del algoritmo

III.4.1 Procesamiento de extremo delantero

La pila de cuatro campos etiquetados Y', C_b' y C_r' en la parte superior de la figura III.7, indica un conjunto de cuatro campos consecutivos de una secuencia de imágenes de prueba o de referencia. La primera etapa del procesamiento transforma los datos Y', C_b' y C_r' en tensiones de cañón electrónico R', G' y B'.

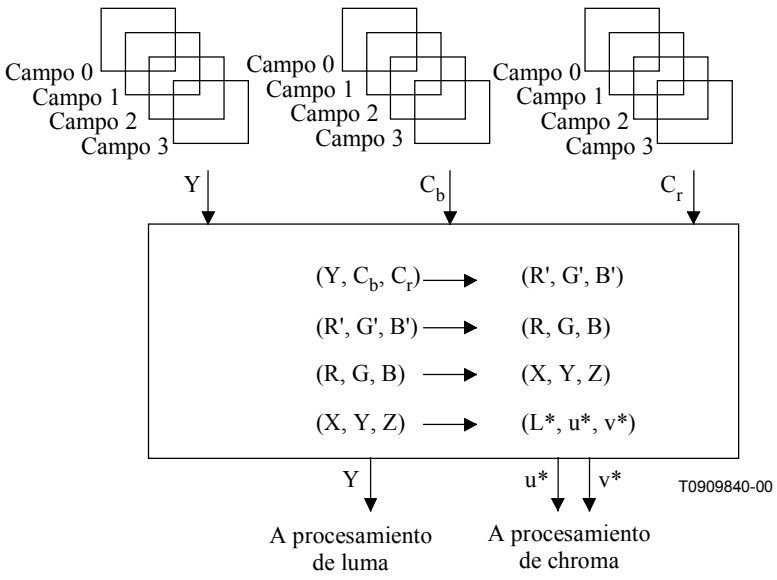


Figura III.7/J.144 – Procesamiento de extremo delantero

La segunda etapa del procesamiento, aplicada a cada imagen R', G' y B', es una no linealidad puntual. En esa etapa se modela el paso de tensiones de cañón electrónico R', G' y B' a intensidades de modelo (R, G y B) de la visualización (fracciones de luminancia máxima). La no linealidad lleva a cabo además un recorte a bajas luminancias en cada plano mediante la visualización.

Tras la no linealidad, se dispone de dos opciones de procesamiento, a saber, de media altura y de altura total. En el caso de exploraciones entrelazadas, las imágenes[1] de media altura se procesan tal como se dan, sin espacios en blanco entre líneas. El modelado de altura total está disponible para exploraciones progresivas (en las que un campo contiene una trama, es decir, una sola imagen en vez de dos campos entrelazados).

A continuación, el vector (R, G y B) de cada píxel del campo es sometido a una transformación lineal (que depende de los luminóforos de la visualización) en coordenadas tricromáticas (X, Y, Z) de CIE 1931. El componente luminancia Y de este valor pasa a procesamiento de luma.

Para asegurar (en cada píxel) la uniformidad perceptiva aproximada del espacio de color con diferencias de color de igual luminancia, se establece la correspondencia entre cada uno de los píxels y CIELUV, norma internacional de espacio de color uniforme (véase Wyszecki y Stiles, 1982). Los componentes de chroma, u* y v*, de este espacio pasan a las etapas de procesamiento de chroma del modelo[2].

III.4.2 Procesamiento de luma

Como se muestra en la figura III.8, a cada valor de luma se le aplica en primer lugar una no linealidad compresiva. A continuación, cada campo de luma es filtrado y muestreado en sentido descendente en una pirámide Gaussiana de cuatro niveles, para modelar la descomposición observada sicofísicamente y sicológicamente de las señales visuales entrantes en diferentes bandas de frecuencia espacial. Tras esa descomposición, el grueso del procesamiento subsiguiente por el modelo consiste en operaciones similares (por ejemplo, filtrado orientado) efectuadas en cada nivel de la pirámide.

---

[1] Las filas de una imagen de media altura corresponden a un campo, es decir, a las líneas pares o las líneas impares de una trama.

[2] El canal de luminancia L* de CIELUV no se utiliza en el procesamiento de luma, sino que es sustituido por una no linealidad visual para la que se ha calibrado el modelo con una gama de valores de luminancia. No obstante, L* se utiliza en el procesamiento de chroma, no obstante, para crear una medida de chroma que sea aproximadamente uniforme y bien conocida por los técnicos de imagen.

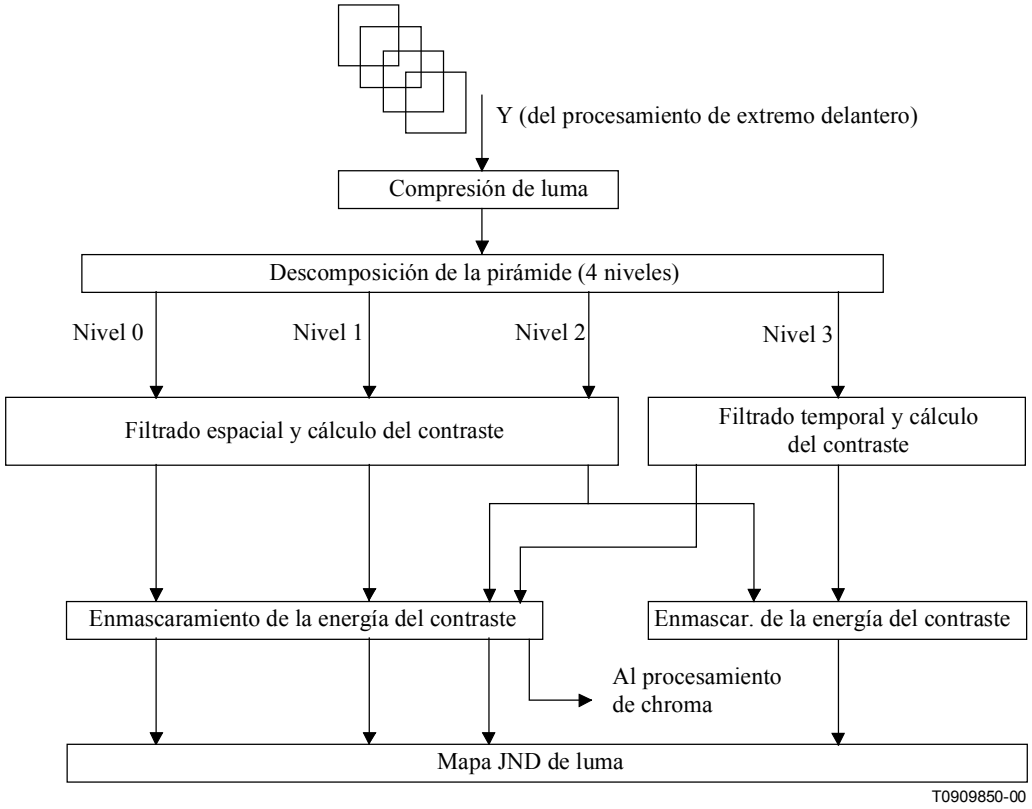


Figura III.8/J.144 – Visión general del procesamiento de luma

Después de este proceso piramidal, la imagen de la pirámide de menor resolución es sometida a un filtrado temporal y se calcula su contraste, y los otros tres niveles son sometidos a un filtrado espacial y se calcula asimismo su contraste. En cada caso, el contraste es una diferencia local de valores de píxels divididos por una suma local, de escala apropiada. Así se establece inicialmente la definición de 1 JND, que pasa a las etapas subsiguientes del modelo[3]. (Mediante calibración iterativa se revisa la interpretación de 1 JND en las etapas intermedias del modelo.) El valor absoluto de la respuesta de contraste pasa a la etapa siguiente, y se mantiene el signo algebraico para reincorporarlo justo antes de la comparación de imagen (cálculo del mapa JND).

La etapa siguiente (enmascaramiento del contraste) es una operación de fijación de ganancia en la que cada respuesta de contraste se divide por una función de todas las respuestas de contraste. Esta atenuación combinada de cada respuesta por otras respuestas locales tiene por objeto modelar los efectos del "enmascaramiento" visual tales como la disminución de la sensibilidad a las distorsiones en zonas de imagen "ocupadas". En esta etapa del modelo, se hace que la estructura temporal (centelleo) enmascare las diferencias espaciales, y además que la estructura espacial enmascare las diferencias temporales. El enmascaramiento de luma se aplica también al lado chroma, como se verá más abajo.

Las respuestas de contraste enmascarado (junto con los signos del contraste) son utilizados para producir el mapa JND de luma. Se hace de la siguiente manera:

- descomponiendo cada imagen en componentes positivos y negativos (rectificación de media onda);
- efectuando un agrupamiento local (promediación y muestreo descendente, para modelar la agregación espacial local observada en los experimentos sicofísicos);

---

[3] La asociación de un contraste constante con 1 JND es una implementación de lo que se conoce como ley de Weber para la visión.

- evaluación de las diferencias de imagen absolutas canal por canal;
- muestreo ascendente hasta el mismo grado de resolución (que será la mitad de la resolución de la imagen original debido a la etapa de agrupamiento);
- evaluación de la norma Q de Minkowski en todos los canales.

III.4.3 Procesamiento de chroma

El procesamiento de chroma sigue las pautas del procesamiento de luma de diversas maneras. Las diferencias de chroma dentro de la imagen (u* y v*) del espacio CIELUV se utilizan para definir los umbrales de detección del modelo de chroma, de forma análoga a cómo se utiliza el contraste (y la ley de Weber) para definir el umbral de detección en el modelo de luminancia. También en analogía con el modelo de luminancia, los "contrastes" cromáticos definidos por las diferencias u* y v* son sometidos a una etapa de enmascaramiento. Una no linealidad de transductor hace que la discriminación de un incremento del contraste entre una imagen y otra dependa de la respuesta de contraste que es común a ambas imágenes.

La figura III.9 muestra que, como en el procesamiento de luma, cada componente de chroma, u* y v*, es sometido a una descomposición piramidal. Sin embargo, mientras que en el procesamiento de luma sólo se necesitan cuatro niveles de pirámide, en el procesamiento de chroma hay siete niveles. Esto refleja el dato empírico de que los canales cromáticos son sensibles a frecuencias espaciales mucho más bajas que los canales de luma (Mullen, 1985). También tiene en cuenta la evidencia de que las diferencias de color se pueden observar en regiones grandes y uniformes.

Para poner de manifiesto la insensibilidad inherente de los canales de chroma al centelleo, se lleva a cabo un procesamiento temporal obteniendo el valor medio de cuatro campos de imagen.

A continuación se efectúa un filtrado espacial mediante una función núcleo de Laplace en u* y v*. Esta operación genera una diferencia de color en u* y v* que (por definición del espacio de color uniforme) está métricamente conectada con las diferencias de color apenas perceptibles. En esta etapa se supone un valor de 1 para significar que se ha alcanzado una sola JND, en analogía con la función del contraste basada en la ley de Weber en el canal de luma. (Como en el caso de luma, la unidad de chroma 1 JND debe ser objeto de una reinterpretación durante la calibración.)

El valor absoluto y ponderado de esta diferencia de color se pasa (con el signo algebraico del contraste) a la etapa de enmascaramiento del contraste. La etapa de enmascaramiento efectúa la misma función que en el modelo de luma. Es algo más sencilla, ya que recibe entrada solamente de los canales de luma y del canal de chroma cuya diferencia se evalúa. Por último, las respuestas de contraste enmascarado se procesan exactamente igual que en el modelo de luma (véase el último párrafo de III.4.2).

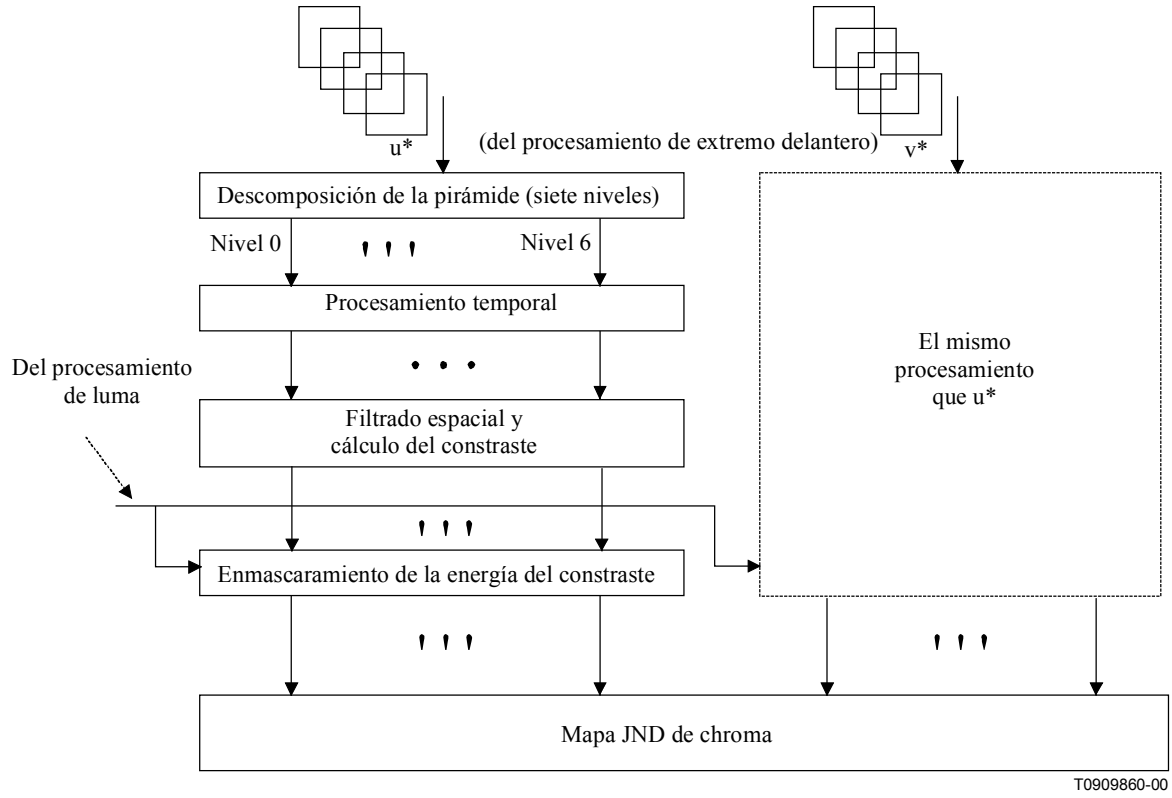


Figura III.9/J.144 – Visión general del procesamiento de chroma

III.4.4 Resúmenes de salida

Para cada campo de la comparación de la secuencia de vídeo, se combinan primero los mapas JND de luma y chroma a fin de obtener un mapa JND total. Ese mapa JND total se calcula como una combinación lineal de la suma y el máximo de los valores de los mapas de luma y chroma, píxel por píxel.

A continuación, cada uno de los tres mapas JND (luma, chroma y luma-chroma combinados) se reducen a un resumen de un solo número, llamado valor PQR (índice de calidad de imagen). Los resúmenes de un solo número se calculan aplicando la norma Q de Minkowski. En este procedimiento, cada valor de píxel de mapa JND se eleva a la Q-ésima potencia. A continuación se calcula el PQR como la Q-ésima raíz de la suma normalizada de todos los valores de píxels elevados a la Q-ésima potencia.

Seguidamente, se calculan tres medidas únicas de la calidad con numerosos campos de una secuencia de vídeo (una para luma, otra para chroma y otra para la combinación luma-chroma). Los valores PQR de cada campo de una secuencia se reducen a un índice de calidad de imagen para toda la secuencia, aplicando de nuevo la norma Q de Minkowski.

III.5 Correlación con resultados objetivos

III.5.1 Visión general

El IRT (Institut für Rundfunktechnik GmbH, de Múnich, Alemania) y Tektronix han concluido recientemente la fase inicial de la investigación sobre la idoneidad de un método de evaluación del índice de calidad de imagen (PQR, *picture quality rating*) objetivo basado en el modelo de visión humana desarrollado conjuntamente por Sarnoff/Tektronix. En esta cláusula se presenta un breve resumen de los resultados de una prueba ciega comparando la medida del índice de calidad de imagen (PQR) y las notas medias de opinión (MOS, *mean opinion score*) de los observadores. El

conjunto de datos de las 60 escenas de vídeo utilizadas en el experimento lo obtuvo el IRT a partir de cinco secuencias de vídeo diferentes pasadas a través de dos codificadores MPEG-2 diferentes a las velocidades comprimidas de 2, 3, 4,5, 7 y 10 Mbit/s. Las MOS fueron determinadas por el IRT y las evaluaciones de PQR objetivas fueron determinadas por Tektronix. En el procedimiento de calificación subjetiva participaron grupos de 25 evaluadores y se siguió de manera estricta el método de UIT-R BT.500-7 [método de escala de calidad continua de doble estímulo (DSCQS)]. Los PQR objetivos los calculó Tektronix con el modelo de visión humana de Sarnoff/Tektronix basado en los principios de la diferencia apenas perceptible. No se ajustó ningún parámetro del modelo a fin de atenerse al conjunto de datos del IRT. Para evitar cualquier posible sesgo en el experimento, Tektronix y el IRT intercambiaron los índices subjetivos y objetivos sólo después de que cada grupo hubo concluido su calificación. Dada la ausencia de ajuste alguno de los parámetros del modelo, que se basan en el conocimiento científico de la visión humana, la concordancia entre resultados subjetivos y objetivos muestra una elevada correlación de 0,88. La correlación con la calidad de la radiodifusión típica es de 0,91. Los resultados se indican en la figura III.12, y son unos resultados prometedores por lo que se refiere a la futura utilización de métodos objetivos en la caracterización y supervisión de la calidad de imagen de vídeo.

III.5.2 Configuración de las pruebas de vídeo y procesamiento

Las escenas de las pruebas de vídeo fueron suministradas por el IRT a Tektronix en formato SMPTE 125M 422-625/50 Hz (es decir, formato de cinta PAL D1). Cada escena tiene una duración de 9 segundos. En lo que sigue, HRC significa "circuito ficticio de referencia" (según la definición ANSI T1A1.5). Antes de pasar el vídeo por los HRC, Tektronix añadió un código de barras cerca de la parte superior de cada trama de vídeo. Ese código se utiliza para determinar si la alineación, horizontal y vertical de los píxels es o no correcta, contar las tramas y tener en cuenta otros factores. Las franjas se cubrieron para las pruebas subjetivas, pero los resultados de una prueba con un pequeño grupo de control y franjas visibles mostraron que las franjas influían poco en las evaluaciones de los observadores. Una vez añadidas las franjas de alineación, el IRT pasó las secuencias por los HRC. Se emplearon dos codificadores de vídeo (IRT[4] y Thomson) a velocidades binarias de 2,0, 3,0, 4,5, 7,0 y 10,0 Mbit/s. Aunque es poco probable que los sistemas de radiodifusión comercial funcionen por debajo de 3 Mbit/s, se incluyeron las escenas a 2,0 Mbit/s para explorar la calidad de funcionamiento más allá de los límites normales. La configuración final de circuitos ficticios de referencia (HRC) consistió en una etapa de conversión PAL seguida de los dos mismos codificadores funcionando a 3 Mbit/s. Previsiblemente, la conversión PAL en particular introducirá alguna desalineación en los subpíxels. A continuación se presentan de forma resumida las secuencias originales y su procesamiento para convertirlas en las escenas de prueba.

[4] El "codificador IRT" fue desarrollado por el IRT y varios socios europeos en el marco de los proyectos Eureka 625 VADI, Race HD-SAT y Race DISTIMA.

Secuencias originales	Circuitos ficticios de referencia (HRC)	Velocidades binarias en Mbit/s
1. Barcelona	1 Codificador IRT	2,0
2. Mobile and Calendar	2 "	3,0
3. NDR	3 "	4,5
4. Football (Soccer)	4 "	7,0
5. Flower Garden	5 "	10,0
	6 Codificador Thomson	2,0
	7 "	3,0
	8 "	4,5
	9 "	7,0
	10 "	10,0
	11 PAL + MPEG (Thomson)	3,0
	12 PAL + MPEG (IRT)	3,0
	13 Referencia – sin compresión	

Configuración de prueba total de **60 escenas** = (**5** secuencias) × [(**2** codificadores) × (**5** velocidades binarias) + **2** PAL]

Barcelona: espectacular desfile multicolor en un gran terreno de juego (véase la figura III.13). La cámara se aleja lentamente y hay poco movimiento. El plano de fondo aporta nitidez a la imagen. La secuencia está llena de colorido, tiene poco movimiento y es nítida.

Mobile and Calendar: secuencia de animación familiar utilizada en la comunidad de compresión de vídeo. Contiene dibujos animados de animales en colores, un tren de juguete en movimiento, un balón que rueda y un calendario con detalles de texto. La secuencia está llena de colorido, tiene poco movimiento y es nítida.

NDR: locutora radiofónica delante de un muro de piedra aglomerada. El muro es muy nítido, pero sin mucho color. La cámara se aleja lentamente. El principal reto al que se enfrenta la compresión son los detalles del muro de piedra. El contenido de movimiento es muy bajo. La secuencia tiene poco movimiento y es nítida.

Football (Soccer en USA): se está jugando un partido de fútbol; la cámara es de gran abertura. No hay mucha acción cercana. El movimiento se puede calificar de moderado. En la escena original, el vídeo está muy desenfocado durante el primer segundo. La secuencia es de movimiento rápido y nítida.

Flower Garden: esta secuencia es muy utilizada en la comunidad de investigación de la compresión del vídeo. La cámara, en un vehículo abierto, se desplaza a velocidad moderada delante de un coloreado jardín de flores. En el plano de fondo hay un molino de viento en movimiento y varias personas. El jardín y las ramas desnudas de los árboles aportan nitidez. El movimiento aparente se puede calificar de moderado. La secuencia está llena de colorido, tiene poco movimiento y es nítida.

En la figura III.13 se muestra una imagen de trama típica de cada una de las secuencias anteriores.

III.5.3 Evaluación subjetiva

En las pruebas se utilizó el método de escala de calidad continua de doble estímulo (DSCQS, *double stimulus continuous quality scale*) de UIT-R BT.500-7.

La estructura de la presentación constó de las siguientes fases, cuya duración se indica en la figura III.10.

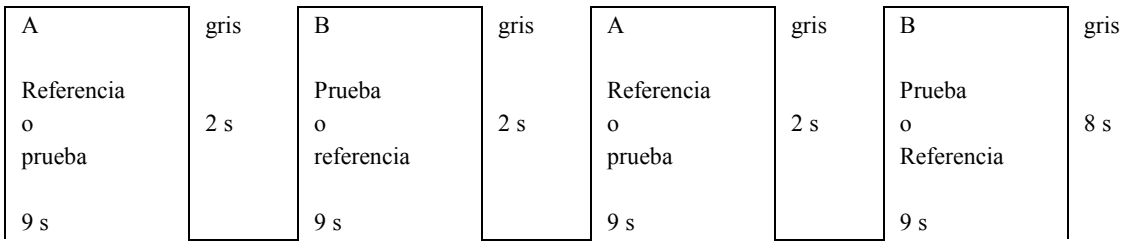


Figura III.10/J.144 – Orden de presentación del método DSCQS

A fue la referencia y B el HRC o viceversa, variando de una prueba a otra. Los evaluadores no conocían el orden. La duración total de cada prueba fue de 50 segundos.

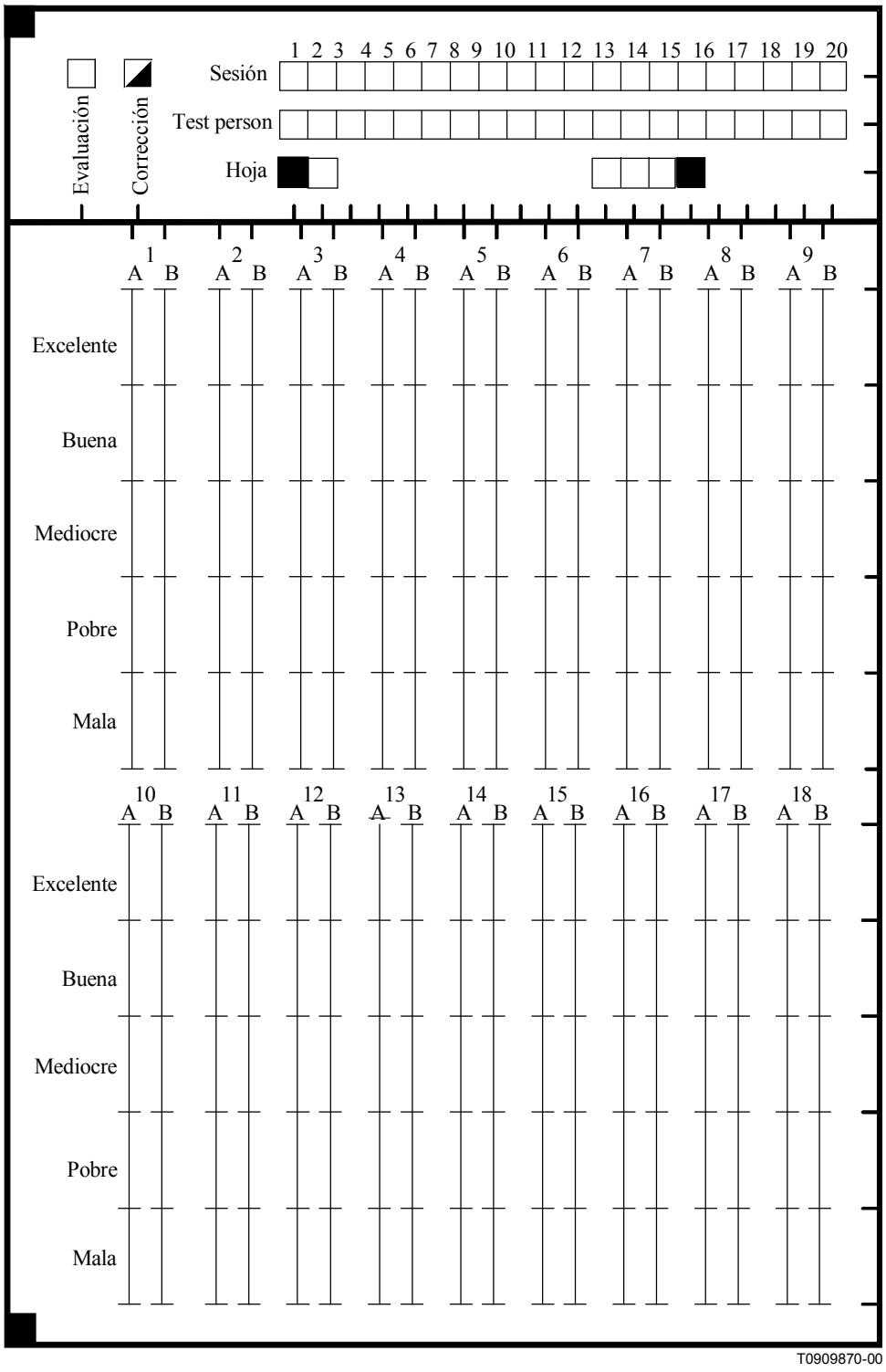


Figura III.11/J.144 – Hoja de prueba utilizada para la evaluación de las secuencias de prueba

Para la evaluación de las secuencias de prueba se utilizó una hoja de prueba del tipo que se muestra en la figura III.11.

La calidad de A y B fue indicada por los evaluadores en una escala lineal. Los términos de calidad en el lado izquierdo significan: excelente, buena, mediocre, pobre, mala. Los resultados se evaluaron electrónicamente y la distancia entre el extremo más bajo de la escala y el indicador de calidad fijado por el evaluador se calculó en cada caso en milímetros. Lo más importante fue la diferencia entre los resultados para referencia y HRC.

Además de efectuar la prueba real, se mostraron ejemplos de secuencias y secuencias de entrenamiento. Al comienzo de la primera sesión se exhibieron cuatro ejemplos. Con ellos se explicaba el método de prueba, y abarcaban toda la gama de calidades que cabía esperar. Se dijo a los observadores que no evaluaran las secuencias porque sólo eran ejemplos. En el cuadro III.2 que sigue se indican los ejemplos de secuencias.

Cuadro III.2/J.144 – Ejemplos de secuencias

Número	Secuencia de prueba	Codificador	Velocidad binaria en Mbit/s
1	Zoom on a street	IRT	3
2	Barcelona 2	Thomson	4
3	Zoom on a street	IRT	10
4	Barcelona 2	Thomson	2

"Zoom on a street" es una producción muy conocida de la BBC que muestra una escena callejera en Edimburgo. Barcelona 2 es una escena de la misma producción que "Barcelona", pero es una aproximación de los participantes.

Las secuencias de entrenamiento tenían que ser evaluadas por los participantes que no sabían que los resultados no eran evaluados. En el cuadro III.3 que sigue se indican las secuencias de entrenamiento.

Cuadro III.3/J.144 – Secuencias de entrenamiento

Número	Secuencia de prueba	Codificador	Velocidad binaria en Mbit/s
1	Renata	Thomson	2
2	Table Tennis	IRT	10
3	Renata	Thomson	4
4	Table Tennis	IRT	2
5	Renata	Thomson	10
6	Table Tennis	IRT	4

"Renata" y "Table Tennis" son secuencias de prueba muy conocidas.

Las sesiones de prueba se estructuraron de la manera siguiente:

Sesión 1: ejemplos (4) – secuencias de entrenamiento (6) – pruebas reales (31)

Sesión 2: secuencias de entrenamiento (6) – pruebas reales (34)

La duración total de la sesión 1 fue de 34 minutos y 10 segundos, el tiempo correspondiente de la sesión 2 fue de 33 minutos y 20 segundos. Veinticinco evaluadores tomaron parte en la serie de pruebas, siendo 15 de ellos personas "externas" (amas de casa, estudiantes, etc.) mientras que las otras 10 eran miembros del personal del IRT (no expertos). La distancia de observación fue de 6 H (H: altura de la imagen). Todas las condiciones eran conformes a UIT-R BT.500-7. Se utilizaron monitores Sony.

Las franjas del código de barras de la parte superior de cada imagen se cubrieron con papel oscuro pegado a la pantalla. Una prueba con un pequeño grupo de cinco evaluadores (personal del IRT, no experto) efectuada cuando las franjas no estaban cubiertas mostró que esa condición no tenía una influencia significativa en los resultados.

Los resultados clave de la prueba subjetiva fueron los valores medios [notas medias de opinión (MOS)] y los intervalos de confianza del 95% de las diferencias entre los resultados para la referencia y el HRC. Puesto que la escala total tiene una longitud de 100 mm, el resultado peor es 100 y el mejor es 0. Un resultado de 20 corresponde a la diferencia entre "excelente" y "buena" o entre "buena" y "mediocre", etc.

III.5.4 Evaluación de la calidad de imagen objetiva

Una vez procesadas las secuencias de vídeo por el IRT a través de los HRC, descritos más arriba, para producir la configuración de prueba, se llevaron a cabo en Tektronix las evaluaciones objetivas del índice de calidad de imagen (PQR). El proceso es como se describe brevemente a continuación:

- se adquiere vídeo a partir de la cinta D1 para computar ficheros de procesamiento digital;
- se aplican algoritmos de alineación temporal y espacial para determinar los desajustes;
- a continuación se alinea de nuevo el vídeo temporal y espacialmente. Para este conjunto de datos, la realineación espacial se efectuó solamente en la ubicación del píxel entero más cercano, por lo que no se invocaron filtros de interpolación. La alineación temporal se hace mediante desplazamiento de trama y no modifica los datos en manera alguna;
- a continuación el vídeo se procesó aplicando el método del índice de calidad de imagen (PQR) objetivo de Sarnoff/Tektronix. Este análisis se llevó a cabo mediante una versión del programa informático del modelo de calidad ejecutado en una estación de trabajo SUN Sparc. El método genera un historial de la calidad de imagen trama por trama de toda la duración del vídeo que permite analizar la calidad de forma continua. Para efectuar una comparación con las evaluaciones subjetivas, esos historiales se resumieron en un índice de calidad de imagen (PQR) global de cada escena, que era una medida de la calidad global a lo largo de la escena.

III.5.5 Comparación de evaluaciones subjetivas y objetivas

La figura III.12 muestra las MOS subjetivas determinadas por el IRT y los PQR objetivos estimados por Tektronix. Las barras de error verticales muestran los intervalos de confianza del 95% para la gama de índices de evaluación de los observadores subjetivos. La relación entre evaluaciones subjetivas y objetivas muestra un buen comportamiento y es monotónica con una elevada correlación de 0,88. A partir de la curvatura hacia la derecha de la relación puede verse que hay una compresión en la evaluación de la calidad de la imagen del observador a medida que la calidad se degrada hacia muy pobre. Se trata de un efecto muy bien conocido en el campo de las pruebas subjetivas, y es coherente con los efectos de la compresión que se han constatado en otras áreas de la percepción humana, tales como las de la sonoridad y el brillo. El grupo de tres puntos del ángulo superior derecho contiene escenas en las que o bien el codificador falló absolutamente en regiones de la escena o bien la calidad era muy pobre. Si se excluyen esos puntos, el coeficiente de correlación aumenta a 0,91. Dado que los índices de calidad objetivos no requerían ninguna adaptación de los parámetros al conjunto de datos de prueba, ni tampoco su optimización, los resultados son muy estimulantes en el sentido de que los métodos objetivos van a contribuir a que se reduzcan el tiempo, el coste y los posibles sesgos asociados a la caracterización y la supervisión del vídeo.

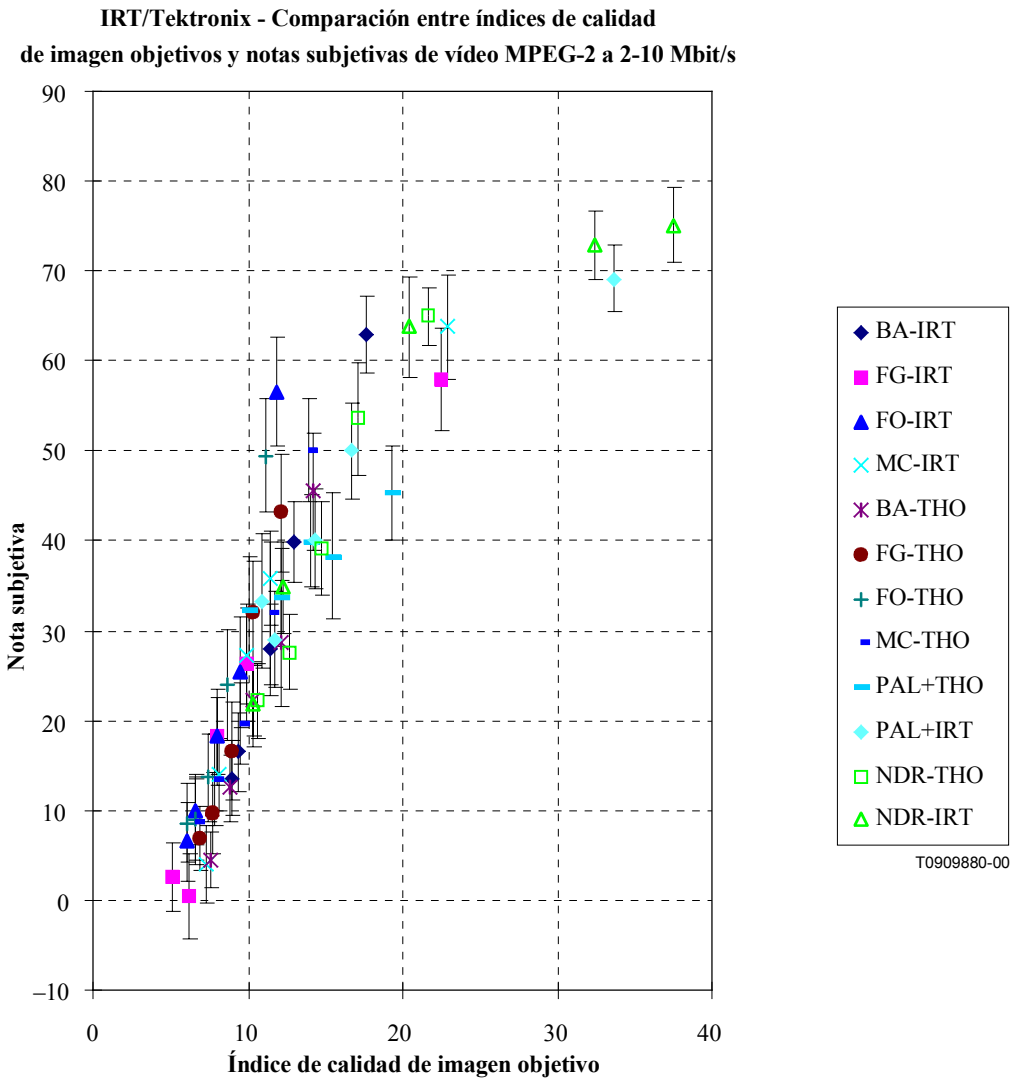
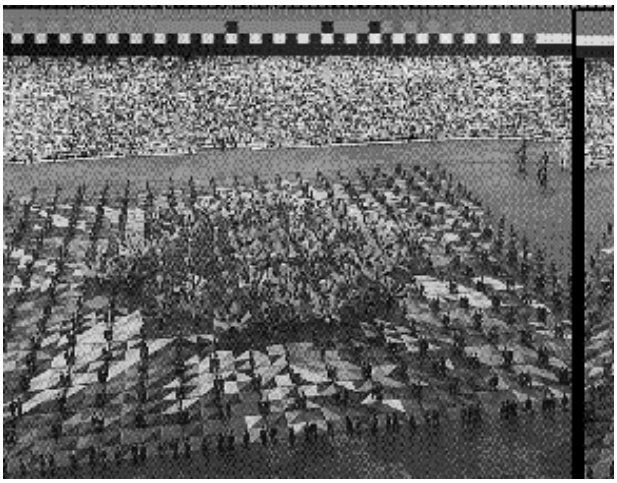
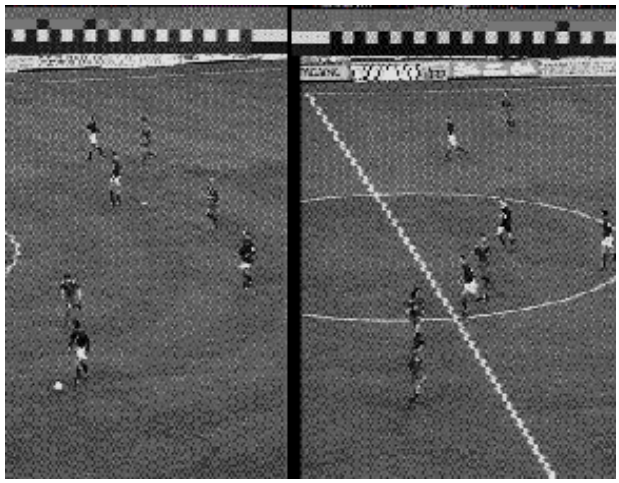


Figura III.12/J.144 – Comparación entre notas medias de opinión subjetivas (MOS) del IRT e índices de calidad de imagen (PQR) objetivos de Tektronix para escenas de prueba MPEG-2 y PAL a 2-10 Mbit/s

Los intervalos de confianza del 95% de las notas subjetivas se indican mediante barras verticales. La correlación entre índices objetivos y subjetivos es de 0,88 para todo el conjunto de datos, y la compresión del observador en el índice de calidad es evidente en las escenas de calidad más deficiente del lado superior derecho. La correlación es de 0,91 si se excluyen las escenas de los datos situados más a la derecha de calidad muy baja.

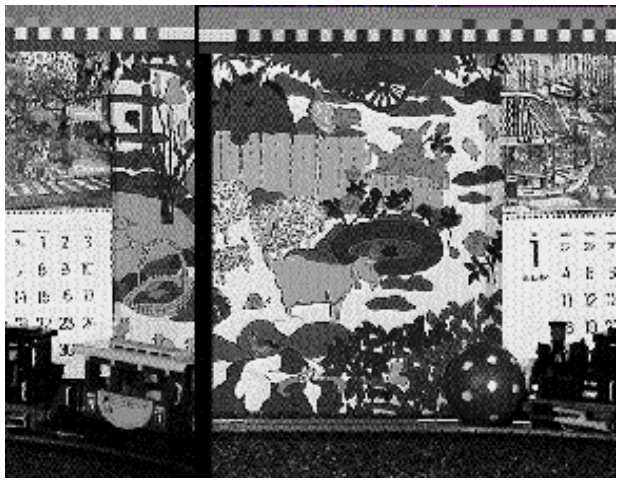


Barcelona



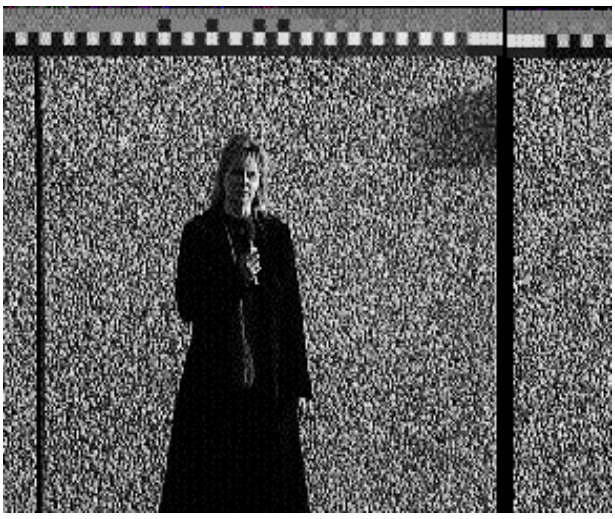
Football

Flower garden



Mobile and calendar

T0909890-00



NDR

Figura III.13/J.144 – Imágenes típicas de secuencias de prueba de vídeo

III.6 Referencias

– LUBIN (J.): The use of psychophysical data and models in the analysis of display system performance, in A.B. Watson (ed.), *Digital Images and Human Vision*, *MIT Press*, pp. 163-178, 1993.

– LUBIN (J.): A visual system discrimination model for imaging system design and evaluation, in E. Peli (ed.), *Visual Models for Target Detection and Recognition*, *World Scientific Publishers*, 1995.

– MULLEN (K.T.): The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings, *J. Physiol.* 359, 381-400, 1985.

– WYSZECKI (G.), STILES (W.S.): Color Science, 2nd ed., *Wiley*, 1982.

APÉNDICE IV

NHK/Mitsubishi Electric Corp.

Resumen

Se ha desarrollado un sistema de evaluación de la calidad de imágenes comprimidas digitales, en el que la degradación de la imagen se calcula en tiempo real teniendo en cuenta la percepción visual humana. En dicho sistema, se considera la sensibilidad al ruido en las frecuencias espacio-temporales en relación con el brillo de la imagen. Este procedimiento ha mejorado la exactitud de la evaluación de la calidad de imagen con respecto a muchos tipos de degradación.

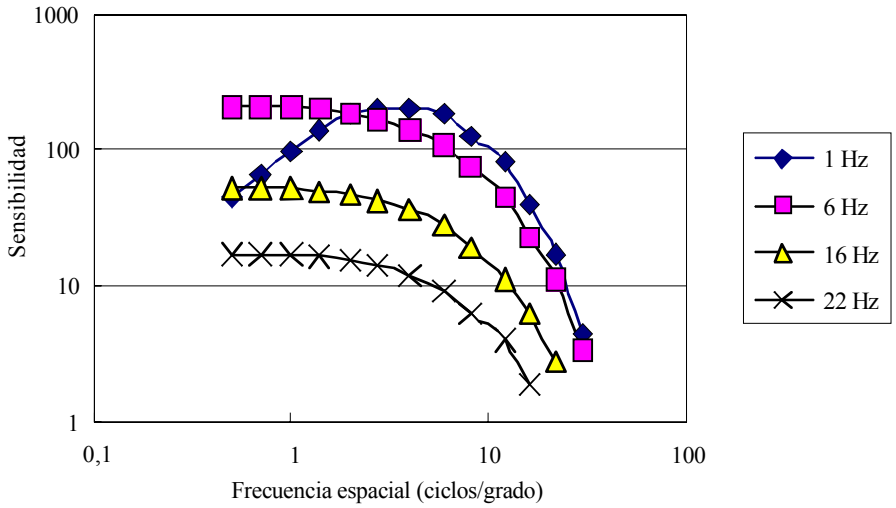
IV.1 Método de evaluación objetiva del deterioro de la calidad

El modelo emula las características visuales humanas utilizando filtros tridimensionales (espacio-temporales), aplicados a las diferencias entre señales fuente y señales procesadas. En la implementación del filtro no se utilizan métodos de análisis de frecuencia tipo bloque, como el de la transformada de coseno discreta (DCT), para evitar posibles efectos mutuos entre los sistemas de codificación y los de evaluación. Las características del filtro se varían en base al nivel de luminancia. La nota de calidad producida se calcula como la suma de las medidas ponderadas obtenidas con los filtros. Con este sistema se pretende evaluar la calidad de imagen en términos de nitidez y repetibilidad, reproduciendo exactamente las funciones visuales en el sistema de evaluación. A continuación se hace una descripción de las características visuales humanas, seguida de explicaciones a propósito de la configuración de los equipos físicos.

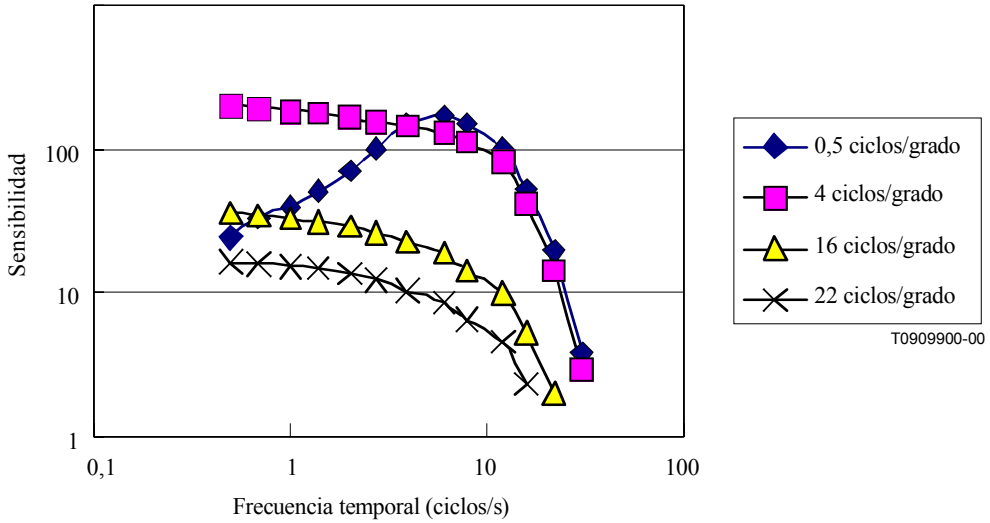
IV.2 Características visuales humanas

IV.2.1 Respuesta en frecuencia espacial de la visibilidad

La respuesta en frecuencia espacio-temporal de la visibilidad humana mostrada en la figura IV.1 ha sido medida por J.G. Robson [1] y otros. La respuesta en frecuencia espacial de la visibilidad muestra una característica seccional de perpendicularidad a la respuesta en frecuencia temporal, alcanzando la simetría rotacional con los ejes ópticos en el centro.



a) Respuesta en frecuencia espacial



b) Respuesta en frecuencia temporal

Figura IV.1/J.144 – Respuesta en frecuencia espacio-temporal de la visibilidad

IV.2.2 Dependencia de la respuesta en frecuencia de la visibilidad con respecto al brillo de la imagen

Por lo que se refiere a la dependencia de la respuesta en frecuencia de la visibilidad con respecto al brillo, las mediciones de Kelly [2] y otros indican que no sólo la respuesta en frecuencia espacial sino también la respuesta en frecuencia temporal dependen del brillo de la imagen. La figura IV.2 muestra la dependencia de la respuesta en frecuencia espacial de la visibilidad con respecto al brillo de la imagen en el caso de una imagen casi fija con una frecuencia temporal de menos de 4 Hz para sensibilidad visual. "td" es la unidad de luminancia de la imagen de fondo del ojo.

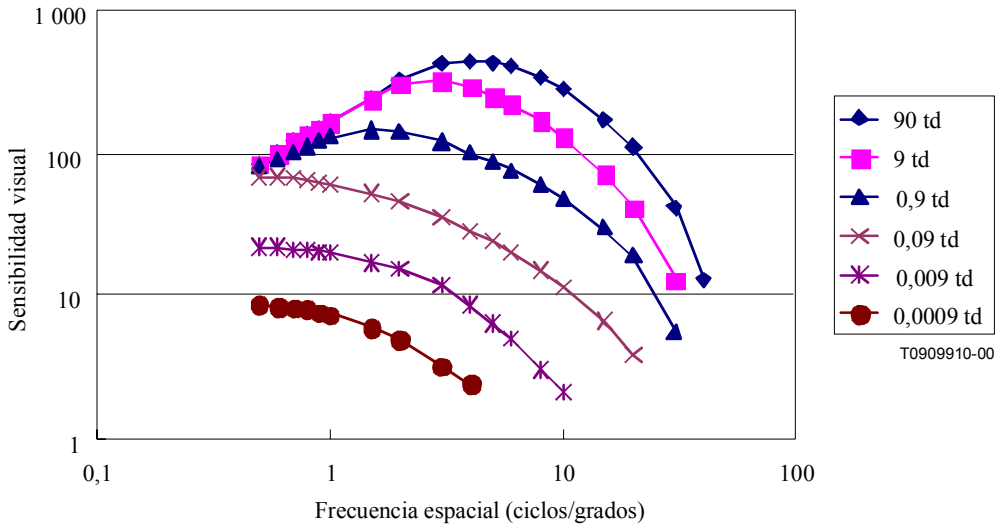


Figura IV.2/J.144 – Dependencia de la respuesta en frecuencia espacial de la visibilidad con respecto al brillo

La figura IV.3 muestra la dependencia de la respuesta en frecuencia temporal de la visibilidad con respecto al brillo en el caso de una imagen uniforme. Los ojos humanos son sensibles normalmente a un centelleo de unos 10 Hz cuando el brillo de la imagen es alto. Cuando es bajo, el centelleo resulta invisible en gran medida.

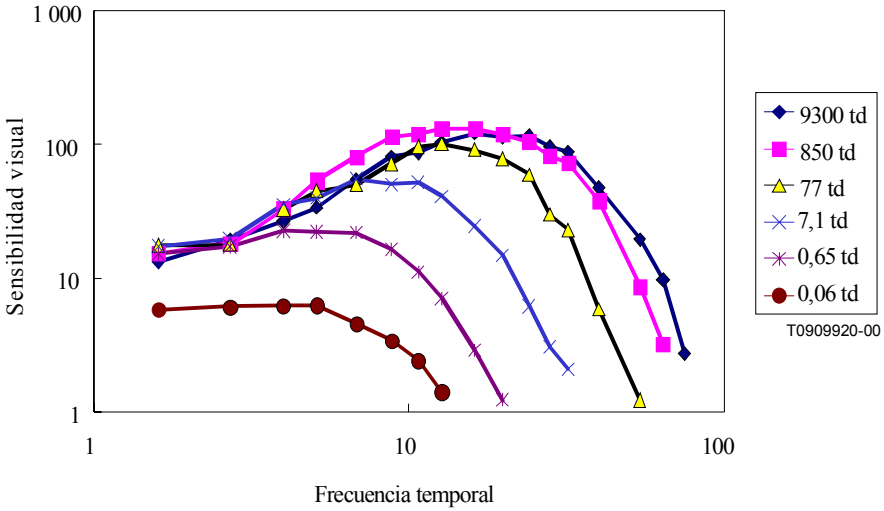


Figura IV.3/J.144 – Dependencia de la respuesta en frecuencia temporal de la visibilidad con respecto al brillo

IV.2.3 Sensibilidad visual dependiente del brillo

La figura IV.4 muestra los límites de la percepción de un ruido aleatorio en el monitor de televisión [3] con diferentes niveles de brillo. Se ha comprobado que existe un cierto grado de dependencia de la sensibilidad visual con respecto al brillo.

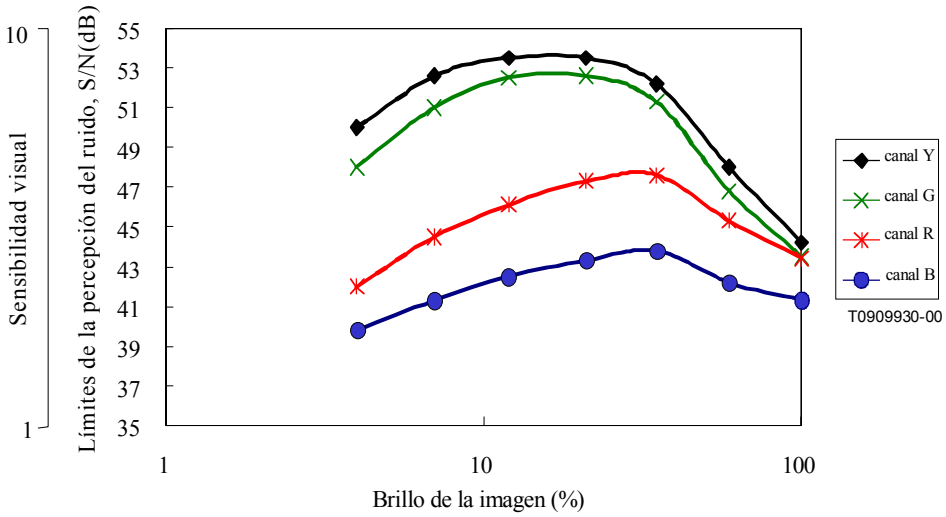


Figura IV.4/J.144 – Límites de la percepción del ruido aleatorio en el monitor de televisión

IV.3 Realización de las funciones visuales por un filtro digital

IV.3.1 Estructura del sistema de evaluación

La figura IV.5 muestra la estructura del sistema de evaluación. En primer lugar, se producen señales de diferencia a partir de las imágenes de la secuencia original y la secuencia de prueba, que se introducen a continuación en el filtro digital tridimensional adaptable al brillo con la misma respuesta en frecuencia tridimensional de la visibilidad y dependencia con respecto al brillo. A continuación, las señales de diferencia filtradas se comparan con la percepción visual en cada píxel. Como resultado de ello, se obtiene una expresión numérica de la distorsión más allá de los límites de la percepción del ojo humano.

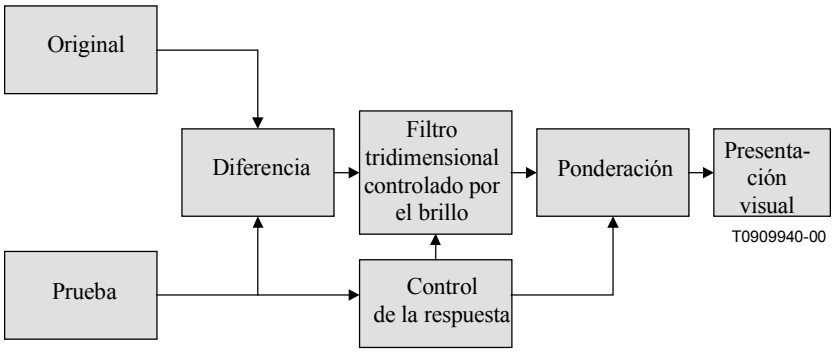


Figura IV.5/J.144 – Estructura de un sistema de evaluación experimental

IV.3.2 Filtro digital tridimensional adaptable al brillo

La figura IV.6 muestra la composición de los filtros digitales tridimensionales, con respuesta en frecuencia y sensibilidad cambiantes de acuerdo con el brillo. Combinando los filtros espaciales y los filtros temporales de acuerdo con el brillo de la imagen, se emula la respuesta en frecuencia de la visibilidad humana.

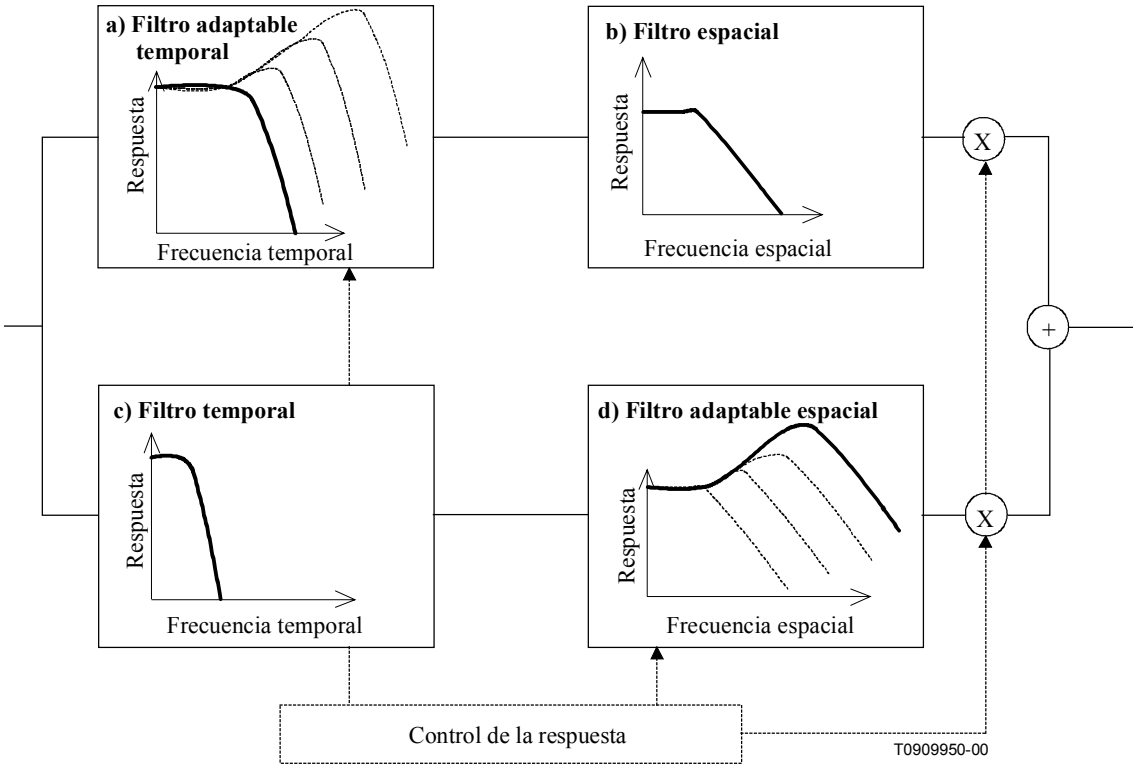


Figura IV.6/J.144 – Composición de filtros digitales tridimensionales, con respuesta en frecuencia y sensibilidad cambiantes de acuerdo con el brillo

IV.3.3 Filtro espacial adaptable dependiente del brillo de la imagen

La figura IV.7 muestra el filtro espacial adaptable d) de la figura IV.6 que se obtiene conmutando entre filtros espaciales con características diferentes.

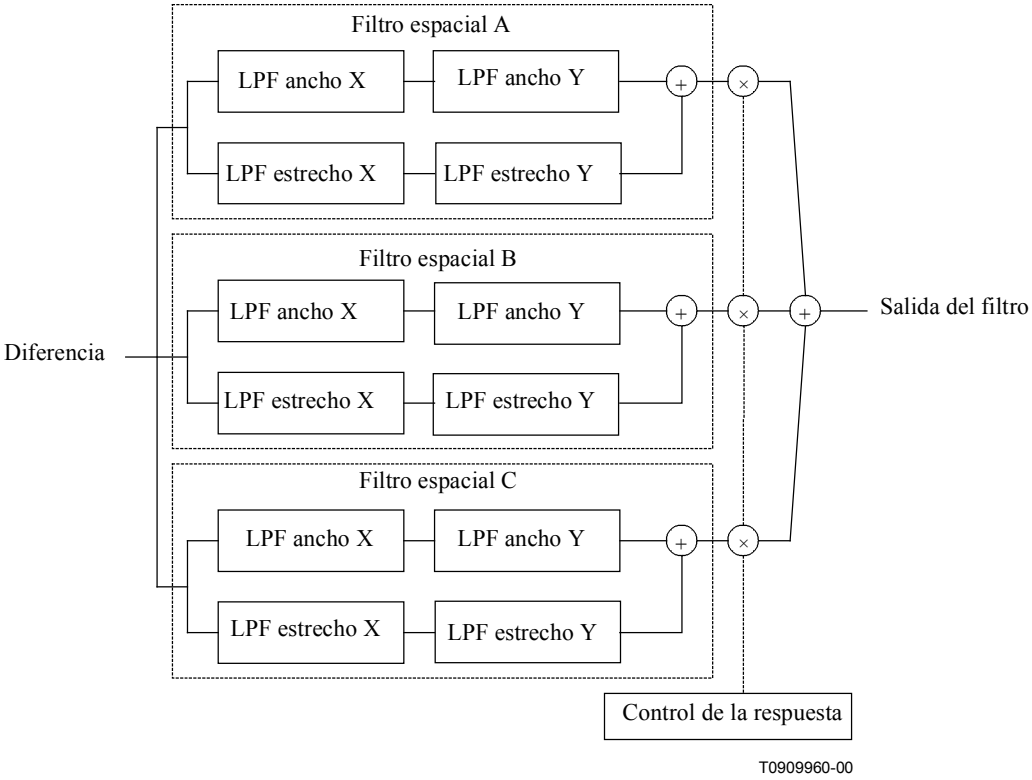


Figura IV.7/J.144 – Filtro espacial adaptable obtenido conmutando entre filtros espaciales

IV.3.4 Respuesta en frecuencia espacial con forma de volcán

Se deduce que las funciones visuales poseen las características de los ejes horizontal y vertical de un filtro espacial de paso de banda. Representando esas características mediante un filtro digital tridimensional se obtiene el perfil en forma de volcán que muestra el gráfico de contornos de la figura IV.8. Dicho perfil representa la respuesta del ojo humano a un deterioro importante en los bordes de la imagen.

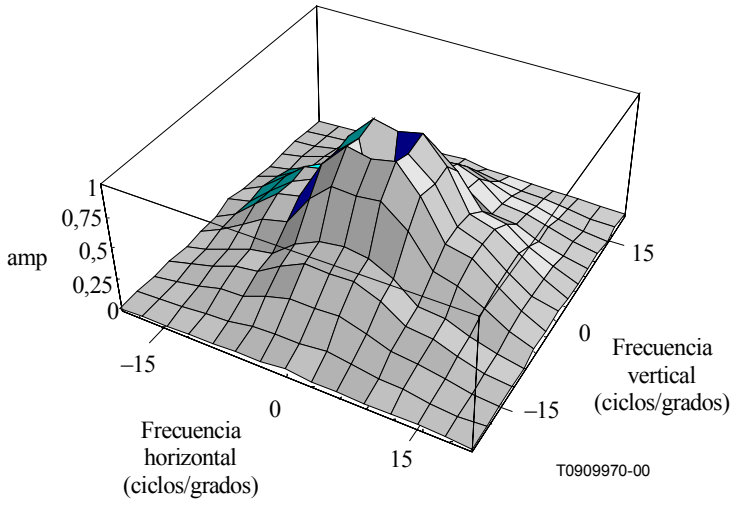


Figura IV.8/J.144 – Filtro espacial con forma de volcán

IV.4 Ejemplo de evaluación mediante el sistema de evaluación de la calidad de imagen

La figura IV.9 ilustra la relación entre las notas o calificaciones de evaluación subjetiva formuladas por 20 expertos en vídeo de acuerdo con UIT-R BT.500 y las notas de evaluación objetiva obtenidas aplicando nuestro nuevo sistema de evaluación. La evaluación la llevamos a cabo utilizando imágenes de componentes y compuestas para la secuencia de prueba, e imágenes de componentes para la secuencia original.

Teniendo en cuenta no sólo la distorsión por la compresión sino también el deterioro de la calidad, incluida la conversión compuesta/componente y los límites de la anchura de banda, puede verse que la calidad de imagen (PQ, *picture quality*) obtenida mediante el sistema de evaluación objetiva tiene una buena concordancia con los resultados del método de escala de calidad continua de doble estímulo (DSCQS) de evaluación subjetiva.

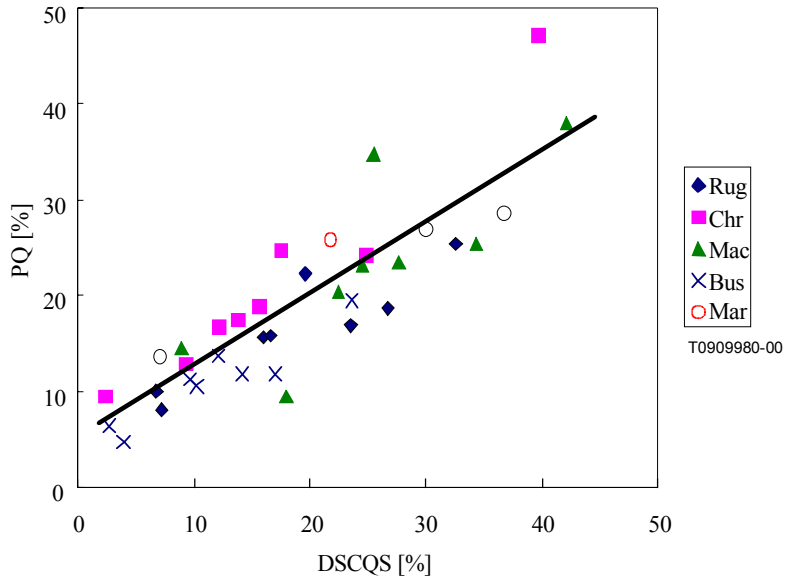


Figura IV.9/J.144 – Relación entre los resultados de la evaluación subjetiva y la evaluación objetiva

La figura IV.10 presenta, para referencia, la relación entre el valor cuadrático medio (RMS) de los errores de las imágenes procesadas y las notas subjetivas. En comparación con la figura IV.9, este gráfico muestra una correlación más baja.

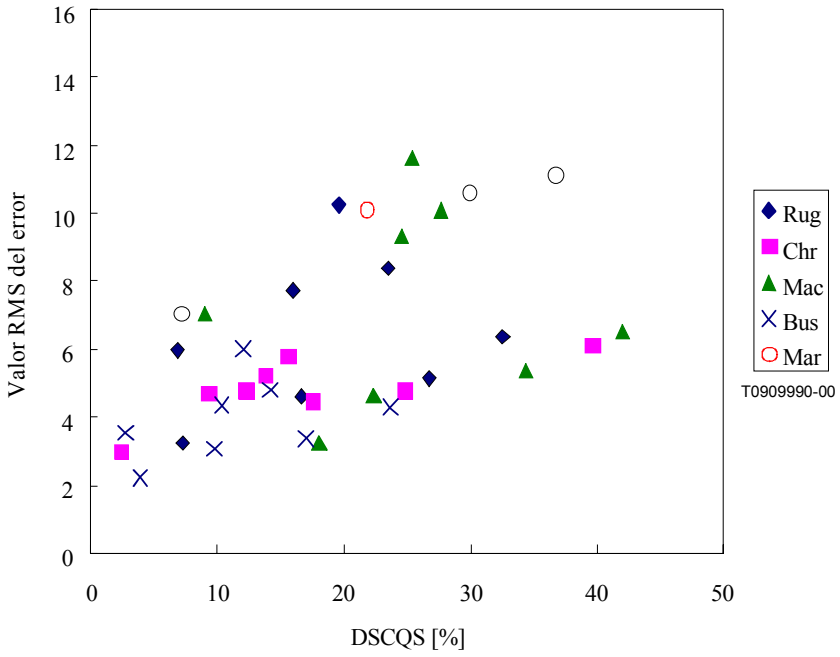


Figura IV.10/J.144 – Relación entre valor RMS del error y notas subjetivas

IV.5 Sistema de evaluación de la calidad de imagen en tiempo real

La figura IV.11 muestra el aspecto externo del dispositivo utilizado en un sistema de evaluación. Las características del sistema son:

1) medición en tiempo real;

2) ajuste automático del retardo del códec y desplazamiento de fase de sincronización; y

3) facilidad de medición porque el sistema tiene las imágenes originales incorporadas.

El sistema se ha mejorado en términos de exactitud mediante la representación fidedigna de las funciones de visibilidad del ojo humano. Normalmente, el ojo humano puede ver un detalle cuando tiene brillo, pero sólo de manera confusa cuando es oscuro. La respuesta en frecuencia espacial varía, obviamente, de acuerdo con el brillo. Además, nuestros ojos pueden ver bien los centelleos de la pantalla cuando brilla. Cuando oscurece, la respuesta en frecuencia temporal varía con la persistencia de la visión. Utilizando el dispositivo del sistema de evaluación, se representan de la mejor manera posible las funciones de visibilidad humana que cambian de manera sustancial de acuerdo con el nivel de brillo.

El sistema de evaluación ha hecho posible efectuar medidas con un alto grado de correlación con la evaluación subjetiva, independientemente del tipo de señal de vídeo, y con una buena representación.

Figura IV.11/J.144 – Aspecto externo del dispositivo utilizado en el sistema de evaluación

IV.6 Referencias

[1] ROBSON (J.G.): Spatial and Temporal Contrast-Sensitivity Functions of the Visual System, *J.Opt. Soc. Am*., pp. 1141-1142, agosto de 1966.

[2] KELLY (D.H.): Visual Responses to Time-Dependent Stimuli. I. Amplitude Sensitivity Measurements, *J. of the Opt. Soc. of Am*., Vol. 51, No. 4, pp. 422-429, abril de 1961.

[3] NISHIDA (Y.), KOIKE (J.), OHTAKE (H.), ABE (M.), YOSHIKAWA (S.): Design Concept for a Low-Noise CCD Image Sensor Based on Subjective Evaluation, *IEEE Trans. ED*., Vol. 36, No. 2, 1989

APÉNDICE V

KDD

Esquema de evaluación de la calidad de vídeo objetiva y evaluación de la calidad de funcionamiento

V.1 Alcance

En los últimos tiempos han empezado a ser de utilización habitual los servicios de radiodifusión y transmisión de televisión digital. En dichos servicios se utilizan códecs de vídeo (dispositivos de codificación de la señal de vídeo) basados en MPEG-2, un método normalizado internacional de compresión de señales de vídeo digitales. Los códecs de vídeo comprenden codificadores, que efectúan la compresión, y decodificadores, que reconstruyen los datos de vídeo comprimidos. Esos dispositivos funcionan eliminando información redundante del enorme volumen de información contenida en las señales de vídeo. Así es posible transmitir la información de manera eficaz utilizando sólo una porción limitada de anchura de banda.

Siempre se produce una cierta degradación de la calidad del vídeo que ha sido comprimido y transmitido utilizando un códec de vídeo. La importancia de la degradación depende del contenido de la imagen. Por lo general, se produce mayor distorsión en las escenas con movimiento rápido, por ejemplo, las de una transmisión deportiva. También hay variaciones en la calidad de la señal de salida generada por códecs diferentes. MPEG-2 es una norma internacional, pero la calidad de tipos específicos de vídeo comprimido sigue dependiendo, en cierta medida, de las características de la función códec implantada por el fabricante en su producto.

En la transmisión televisiva, sobre todo en las clases de vídeo TV1, TV2 y TV3 (contribución, distribución primaria y distribución secundaria) [1], hay que tratar de conseguir una alta calidad supervisando de manera continua la calidad de las imágenes transmitidas.

En la transmisión analógica convencional con modulación de frecuencia, se produce poca degradación de la imagen imputable al contenido o a la modulación analógica, con lo que la calidad permanece estable. Pero en la transmisión de vídeo digital comprimido, la calidad de la imagen varía, como se indica más arriba, de acuerdo con la naturaleza del contenido y el códec empleado y, previsiblemente, la comprobación de la calidad de este tipo de vídeo será una operación sumamente compleja.

Por ello, se piensa que es preciso normalizar un esquema de evaluación de la calidad de imagen de los códecs de vídeo basados en la norma MPEG-2, utilizados principalmente en las clases de vídeo TV1, TV2 y TV3. Se considera que en esas clases son necesarias las funciones siguientes:

- Evaluación genérica para diversos tipos de contenido de vídeo
  Se soportan formatos de vídeo analógico/digital, compuesto/en componentes.
- Evaluación en tiempo real
  Alineación precisa temporal y espacial entre una señal original y una señal de salida de códec.
- Evaluación sensible y exacta de distorsiones sutiles y complejas.

Teniendo en cuenta todo lo anterior, se propone un nuevo esquema de evaluación y su implementación en base a las características de la percepción visual humana, que permite mediciones muy precisas de la calidad vídeo en [2]. En el presente apéndice, se informa de los resultados de la verificación del esquema.

V.2 Esquema de evaluación de la calidad vídeo objetiva

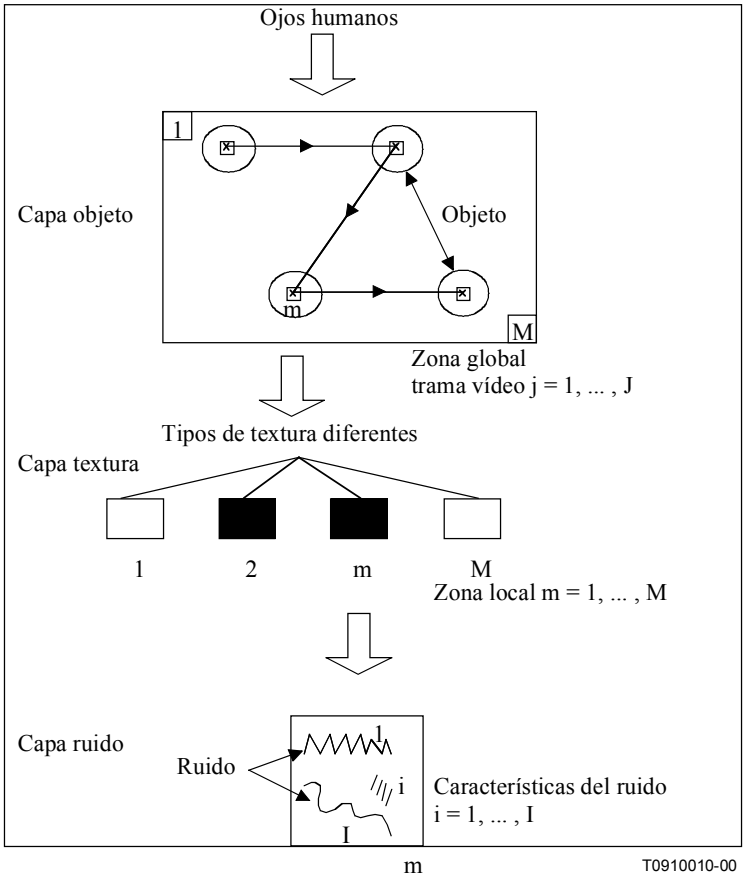


Figura V.1/J.144 – Modelo de tres capas para señal de vídeo

La figura V.1 muestra el modelo de evaluación de la calidad de imagen de tres capas según la visión del ojo humano. Por lo general, el ojo humano no puede ver una trama completa con una sola mirada, sino solamente una zona puntual de una trama, en torno al punto de observación del ojo humano, y reconoce la textura así como la calidad de la zona en función de los grados y características de la mezcla de ruido en esa textura. La trama completa se capta desplazando el punto de observación entre los objetos, que son componentes de imagen de la trama, y al mismo tiempo se efectúa la evaluación de la calidad de imagen para la totalidad de la trama. En este proceso, la calidad de imagen se determina por el ruido a lo largo de una trama. Por consiguiente, para efectuar una medición objetiva de la calidad de imagen subjetiva, se emplean estructuras de imagen de tres capas de macro a micro (capas objeto, textura y ruido), y se propone un esquema de ponderación del ruido de abajo a arriba que utiliza una función de ponderación particular en cada capa teniendo en cuenta la percepción visual humana (figura V.2).

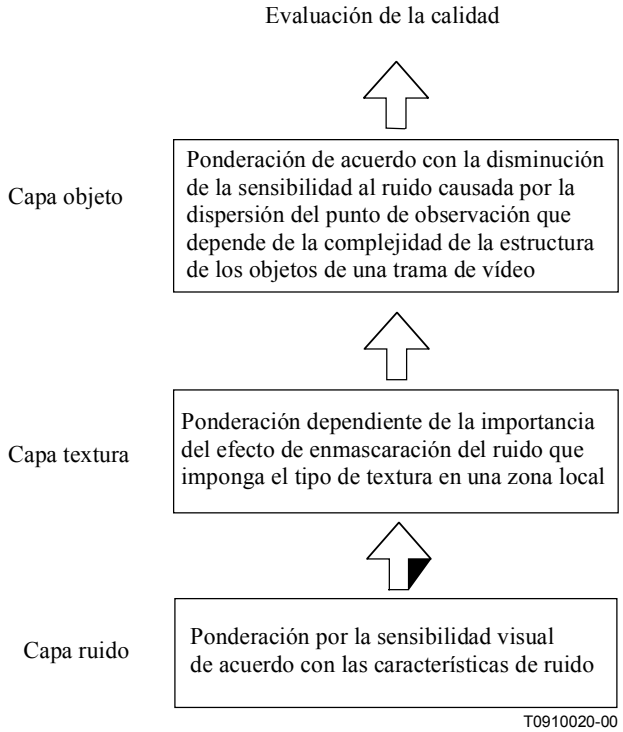


Figura V.2/J.144 – Ponderación del ruido de abajo a arriba con estructura de tres capas

En primer lugar, en la capa ruido, se ponderan los ruidos comunes de un proceso de compresión de vídeo, tales como el ruido de alta frecuencia, el ruido de baja frecuencia, el ruido de chroma, la inestabilidad, el centelleo y así sucesivamente, dependiendo de sus grados y características. Para esta ponderación, conviene efectuar una conversión de frecuencia a fin de clasificar estos ruidos. En segundo lugar, en la capa textura, se clasifican las zonas puntuales locales en varios grupos de acuerdo con sus tipos de textura. Dichos tipos incluyen, por ejemplo, "textura de detalle", tal como un bosque, árboles y un estadio en el que el ruido está fuertemente enmascarado, y "textura plana", tal como la piel de un ser humano, o bien el firmamento, en donde el ruido es fácilmente reconocible. Los ruidos se ponderan más o menos, por consiguiente, de acuerdo con sus tipos de textura. Por último, en la capa objeto, el grado de dispersión del punto de observación se predice midiendo el grado de complejidad de la estructura de los objetos en la trama de vídeo. A continuación se ponderan los ruidos de toda la trama en correspondencia con una disminución de la sensibilidad al ruido causada por la dispersión.

Para obtener las expresiones matemáticas de los procesos de ponderación, se define lo siguiente:

P(j,m,i): Potencia de un ruido i en una zona local m de una trama j

h_i: Función de ponderación de un ruido i

C(j,m): Textura en una zona local (j,m)

t_c: Función de ponderación del ruido en una textura C

G(j): Parámetro que indica el grado de complejidad de la estructura de los objetos de una trama j

qG: Función de ponderación de un ruido dependiente del grado de dispersión de un punto de observación

Teniendo en cuenta esas definiciones, los ruidos se suman en orden desde la capa baja hasta la capa alta.

En la capa ruido, sumando el ruido ponderado mediante la h_i correspondiente a las características del ruido en una zona local (j,m), se calcula $WMSE_{NL}$ como sigue:

$$WMSE_{NL}(j,m) = \frac{1}{I}\sum_{i=1}^{I} h_i \cdot P(j,m,i) \tag{V-1}$$

Seguidamente, en la capa textura, sumando $WMSE_{NL}(j,m)$ en toda la trama (m = 1, ..., M) ponderado mediante la t_c correspondiente a una textura C(j,m) en una zona local (j,m), se calcula $WMSE_{TL}(j)$ como sigue:

$$WMSE_{NL}(j) = \frac{1}{M}\sum_{m=1}^{M} t_c(j,m) \cdot WMSE_{NL}(j,m) \tag{V-2}$$

Por último, en la capa objeto, tomando un valor medio de $WMSE_{TL}$ en las tramas j = 1, ..., J ponderadas mediante el G(j) correspondiente al grado de dispersión del punto de observación, se calcula $WMSE_{OL}$ como sigue:

$$WMSE_{OL} = \frac{1}{J}\sum_{j=1}^{J} q_G(j) \cdot WMSE_{TL}(j) \tag{V-3}$$

A continuación se convierte $WMSE_{OL}$ en WSNR y se calcula la escala de calidad continua de doble estímulo (DSCQS) (0-100%) definida en UIT-R BT.500-7 como sigue:

$$WSMR(dB) = 10\ \log_{10}\frac{255^2}{WMSE} \tag{V-4}$$

$$D(\%) = f(WSNR) \tag{V-5}$$

V.3 Implementación

El dispositivo del sistema consta de dos partes: un módulo de sincronización, que permite efectuar una comparación precisa entre el vídeo reconstruido y el vídeo original, y un módulo de cálculo, que determina la calidad vídeo con referencia a las características de la percepción visual humana. La figura V.3 muestra la configuración del sistema. En el cuadro V.1 se describen los parámetros principales. Como muestra el cuadro V.1, se soportan tanto señales compuestas (NTSC) como componentes con muestreos completos.

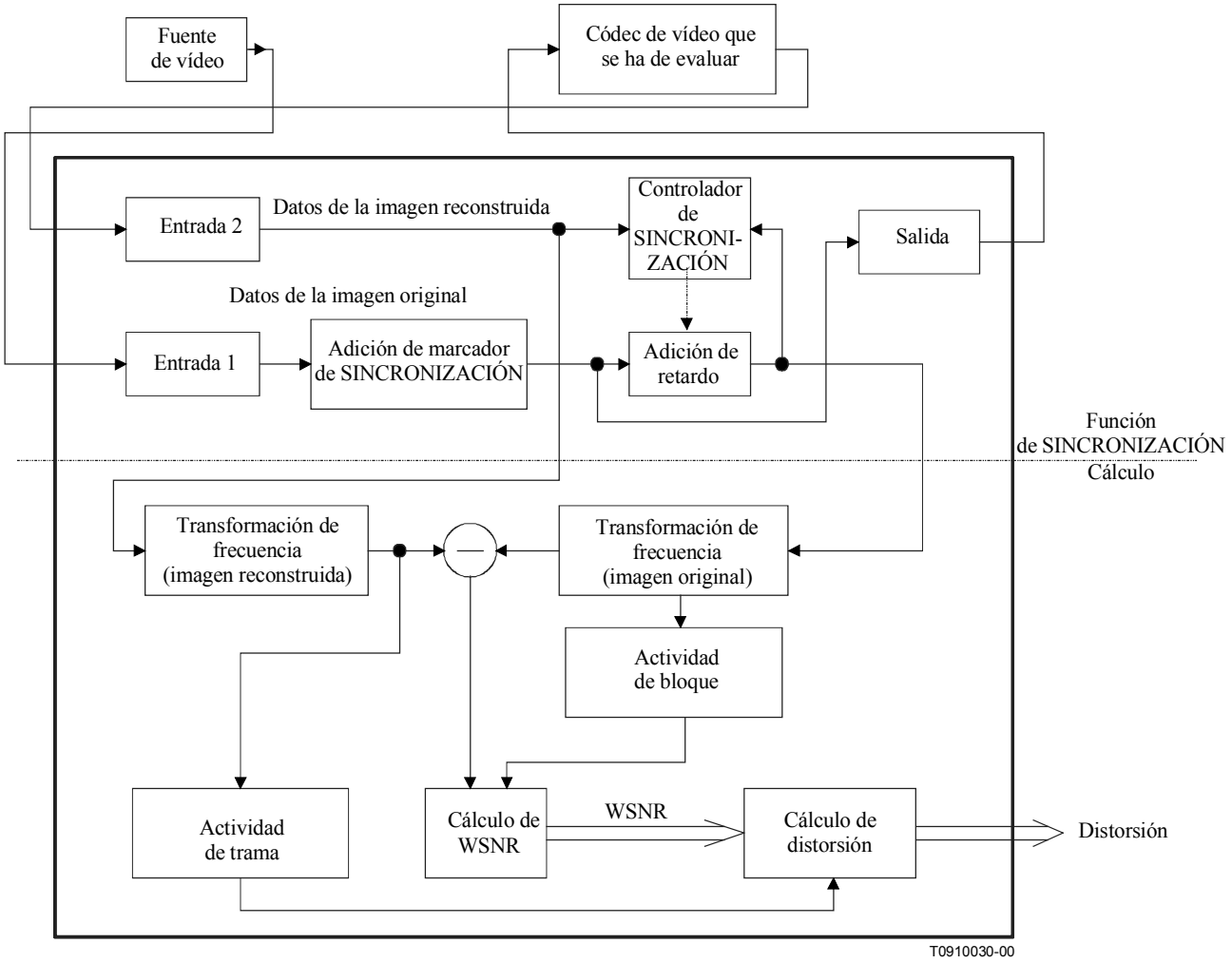


Figura V.3/J.144 – Configuración del sistema

V.3.1 Módulo de sincronización

Las señales de televisión procedentes de la fuente de vídeo original se introducen en el dispositivo del sistema mediante el módulo de entrada 1 y se marcan con un marcador de sincronización que varía con cada trama. A continuación, las tramas con marcadores son enviadas al módulo de retardo, en cuya memoria se almacenan. Al mismo tiempo, las tramas son enviadas a través del módulo de salida al códec de vídeo que se ha de evaluar. El códec de vídeo comprime las tramas, que son introducidas de nuevo en el sistema a través del módulo 2 y comparadas con las tramas marcadas almacenadas en el módulo de retardo del códec de vídeo que se ha de evaluar. Por último, el módulo de sincronización lleva a cabo una alineación temporal (retardo de trama) y una alineación espacial (desplazamiento de línea y píxel) precisas, de modo que el grado de deterioro de la calidad que se describe más abajo concuerde lo más posible con la evaluación subjetiva efectuada por los observadores.

Con estas operaciones se obtiene la sincronización necesaria para la evaluación, y los marcadores utilizados en las mismas se diseñan de tal manera que funcionen correctamente incluso en los procesos con señal fuertemente distorsionada, tales como la compresión intensa, la separación Y/C y los filtrados en códec de vídeo.

V.3.2 Módulo de cálculo

A diferencia de lo que ocurre en la visión humana, el cálculo de la calidad de imagen tiene lugar siguiendo un proceso de abajo a arriba, construyendo el todo a partir de las diversas partes. En

primer lugar, para evaluar el efecto de las variaciones de sensibilidad debidas a las frecuencias espaciales del ruido, se obtiene un ruido valor de la diferencia de los componentes de frecuencia de la imagen original y la imagen reconstruida. Ese valor se introduce en el módulo relación señal/ruido ponderada (WSNR, *weighted signal-to-noise ratio*), que asigna ponderaciones de sensibilidad diferentes a cada región de frecuencia. Al mismo tiempo, obtiene un valor (la actividad de bloque) que indica si cada uno de los bloques de la imagen está vacío u ocupado. El efecto de enmascaramiento del ruido se aplica también para obtener una WSNR global.

Por último, se obtiene un valor que indica el tamaño de los objetos que forman la imagen (la actividad de trama). De esta manera, el sistema puede estimar el grado en que disminuye la sensibilidad al ruido debido a la dispersión de la degradación de la calidad, que se obtiene aplicando la disminución de la sensibilidad al ruido en la WSNR.

Cuadro V.1/J.144 – Parámetros principales

Formato de señal de vídeo aplicables	Señal compuesta NTSC Señal de componentes 525/60 Digital en serie D1
Frecuencia de muestreo (Entrada analógica)	14,318 MHz (NTSC) 13,5 MHz (Componente Y) 6,75 MHz (Componente C)
Códec aplicable	Códec basado en MPEG-1, 2 Códec compuesto, etc.
Zona de evaluación efectiva	768 píxels~480 líneas (NTSC) 720 píxels~480 líneas (Componente Y) 360 píxels~480 líneas (Componente C)
Análisis de la señal	Transformada Hadamard (NTSC) Transformada de coseno discreta (Componente) Alternativa: Transformada de Fourier
Ponderación del ruido	Sensibilidad visual a la frecuencia espacial Efecto de enmascaramiento del ruido Dispersión del punto de observación
Resultado de la evaluación	Evaluación de la calidad de la imagen (Distorsión, %) WSNR (dB) SNR (dB)
Interfaz de la señal de control	RS-232C

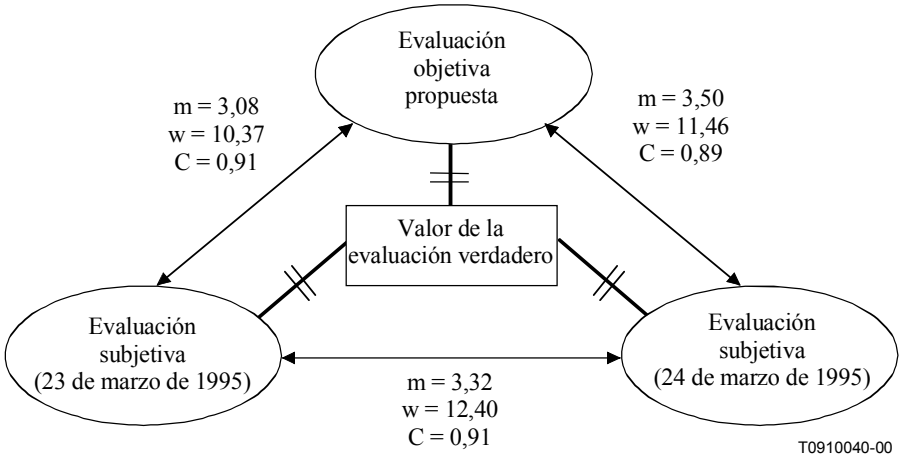
V.4 Resultados de la verificación

Se compararon los resultados de la evaluación del esquema propuesto con los resultados de la prueba de evaluación subjetiva que ya habían sido calificados siguiendo UIT-R BT.500-7. Los objetivos de la evaluación son MPEG-2 SP@ML con 5 Mbit/s, 7 Mbit/s y 10 Mbit/s aplicados para las señales de prueba de televisión de componentes 4:2:2 de UIT-R BT.601. Se trata de 17 datos entre los que figuran Mobile, Flower garden Cheer leaders, etc. Se dispuso por tanto de 17 datos × 3 bits = 51 muestras (cuadro V.2).

Con esas muestras se llevaron a cabo pruebas de evaluación subjetiva durante dos días (23 y 24 de marzo de 1995) en las mismas condiciones y con los mismos observadores. En la figura V.4 se muestran el "triángulo" de los resultados de la evaluación objetiva y de la evaluación subjetiva de dos días.

Cuadro V.2/J.144 – Lista de datos de prueba

1	Susie
2	Popple
3	Table tennis
4	Mobile & Calendar
5	Autumn leaves
6	Football
7	Storm
8	Cheer leaders
9	Cast
10	Cruising
11	Bicycle
12	Horse riding
13	Summer flowers
14	Ferris wheel
15	Flower garden
16	Kiel Harbor 4
17	Balls of wool



m Error cuadrático medio
w Error cuadrático más desfavorable
C Correlación

Figura V.4/J.144 – Comparaciones con pruebas de evaluación subjetiva

La figura V.4 demuestra que la exactitud de la evaluación expresada en rmse, rwse y correlación de los resultados de las tres evaluaciones es casi igual a partir del centro del triángulo, que es el valor verdadero de la evaluación. Además, la figura V.5 presenta las distribuciones de las 51 muestras entre una evaluación objetiva y dos evaluaciones subjetivas. Las muestras de los tres gráficos están distribuidas aleatoriamente, pero puede apreciarse una diferencia sutil en cada distribución. En la distribución de la comparación subjetiva de los días 23 y 24, es uniformemente aleatoria, pero se pueden observar desigualdades en las distribuciones de las comparaciones de la evaluación objetiva y subjetiva dependiendo del rango de las notas de calificación. Esto es, tanto el gráfico del día 23 como el del día 24 correspondientes al esquema objetivo presentan trazados del muestreo con correlación más alta entre 20% y 40% pero menor correlación entre 10% y 20%. Se seguirá estudiando para eliminar este efecto.

De todo lo anterior se concluye que es posible utilizar el esquema propuesto además del de UIT-R BT.500-7.

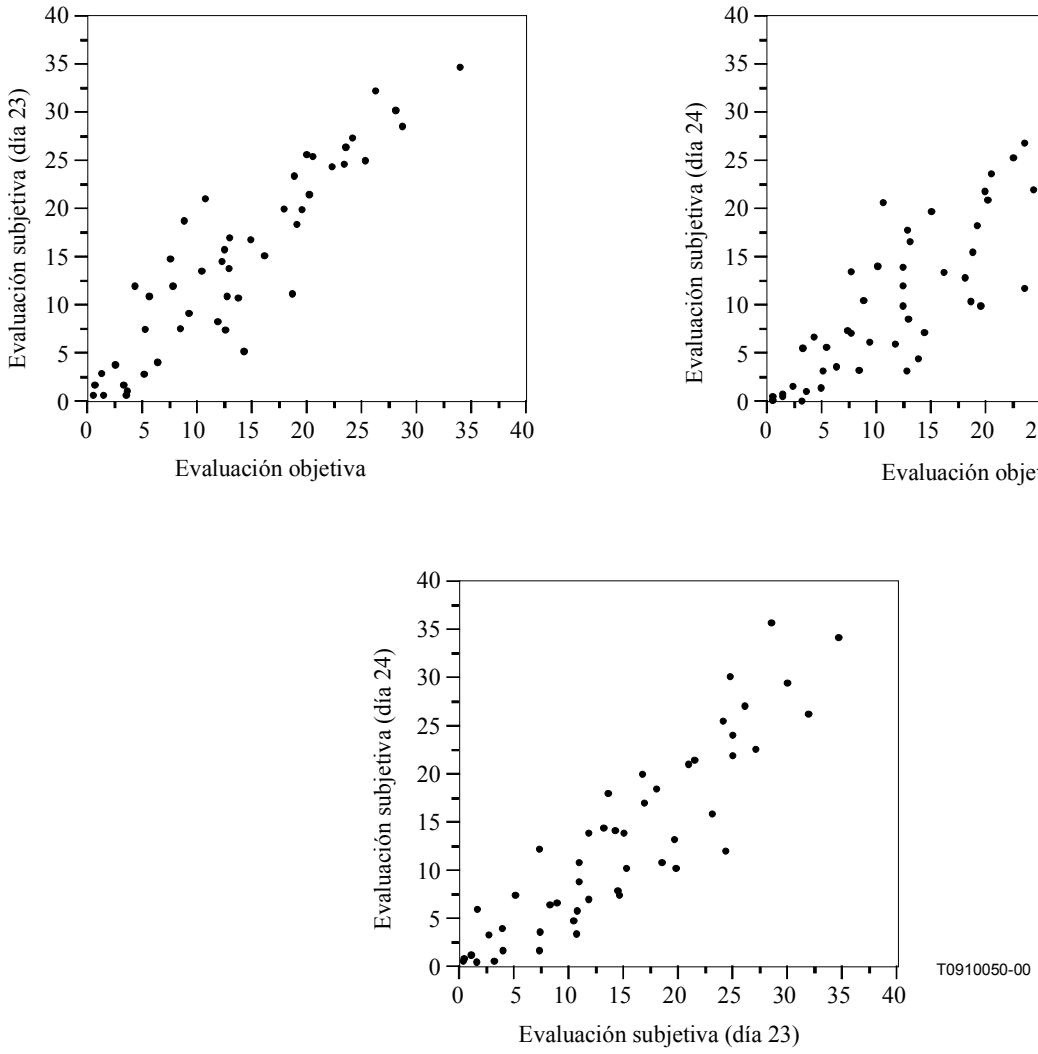


Figura V.5/J.144 – Comparaciones entre una evaluación objetiva y dos evaluaciones subjetivas

V.5 Referencias

[1] *2nd version of Table defining video quality classes*, Expert meeting on subjective and objective video quality assessment, Turín, 14-16 de octubre de 1997.

[2] *Progress report on development of digital compressed picture quality assessment system in Japan*, SG 9 Document D15 Ginebra, 21-25 de abril de 1997.

APÉNDICE VI

EPFL

La medición de la distorsión perceptiva (PDM, *perceptual distorsion metric*) elaborada por EPFL se basa en un modelo espacio-temporal del sistema visual humano. Consta de cuatro etapas, a través de las cuales pasan la secuencia de referencia y la secuencia procesada. La primera etapa convierte la entrada en un espacio de colores antagónicos. En la segunda etapa se lleva a cabo una descomposición perceptiva espacio-temporal en canales visuales independientes de frecuencia temporal, frecuencia espacial y orientación diferentes. La tercera etapa modela los efectos del enmascaramiento del diagrama simulando mecanismos de excitación e inhibición de acuerdo con un modelo de control de ganancia del contraste. La cuarta y última etapa de la medición sirve a modo de etapa de agrupación y detección y en ella se calcula una medida de la distorsión a partir de la diferencia entre las señales de salida del sensor de la secuencia de referencia y la secuencia procesada.

APÉNDICE VII

NASA

VII.1 Introducción

La infraestructura emergente para vídeo digital precisa de un componente fundamental, a saber, un sistema fiable de medición automática de la calidad visual. Dicho sistema es esencial para la evaluación de los códecs, la supervisión de las transmisiones de radiodifusión y para asegurar la máxima eficacia en la compresión de las fuentes y la utilización de las anchuras de banda de comunicación. Este apéndice describe una medida de la calidad vídeo nueva, llamada calidad vídeo digital (DVQ), que se puede utilizar para medir la calidad vídeo de manera automática.

VII.2 La medida DVQ

Todas las medidas de la calidad vídeo son, por su propia naturaleza, modelos de la visión humana. La medida DVQ es un intento de incorporar muchos aspectos de la sensibilidad de la visión humana en un algoritmo de procesamiento de imágenes sencillo. La sencillez es un objetivo importante ya que lo ideal sería que la medición se efectuara en tiempo real y sólo necesitara unos pocos recursos de cálculo. Uno de los elementos más complejos de las otras mediciones propuestas y al que más tiempo hay que dedicar son las operaciones de filtrado espacial efectuadas para implementar los filtros espaciales múltiples de paso de banda característicos de la visión humana. Esta etapa se acelera utilizando la transformada de coseno discreta (DCT) para la descomposición en canales espaciales. Esto representa una ventaja importante ya que se dispone de soportes físicos y lógicos eficaces para efectuar esa transformación, y porque es posible que en muchas aplicaciones, la transformación ya se haya realizado como parte del proceso de compresión.

La figura VII.1 muestra una visión general de las etapas del procesamiento de la medida DVQ. Dichas etapas se describen con más detalle en otro lugar [1] a [3]; aquí sólo se analizan brevemente. Los datos que se introducen en la medida son un par de secuencias de imágenes en color: la de

referencia (R) y la de prueba (T). El primer paso consiste en la realización de diversos muestreos, arrastres de imágenes y transformaciones del color que sirven para limitar el procesamiento a la región que interesa y para expresar las secuencias en un espacio de percepción del color. En esta etapa se efectúa también el desentrelazado y la descorrección gamma del vídeo de entrada. Las secuencias se someten a continuación a una división en bloques (BLK, *blocking*) y una transformada de coseno discreta (DCT), y los resultados se convierten seguidamente en contraste local (LC, *local contrast*). El contraste local es la relación entre la amplitud de la DCT y la amplitud de DC para el bloque correspondiente. El paso siguiente es una operación de filtrado temporal (TF, *temporal filtering*) que implementa la parte temporal de la función de sensibilidad al contraste. Se lleva a cabo mediante un filtro adecuado de segundo orden discreto y recursivo. Los resultados se convierten a continuación en diferencias apenas perceptibles dividiendo cada coeficiente de la DCT por su umbral visual respectivo. Así se implementa la parte espacial de la función de sensibilidad al contraste (CSF, *contrast sensitivity function*). En la etapa siguiente, se sustraen ambas secuencias. La secuencia de diferencias se somete a continuación a una operación de enmascaramiento del contraste (CM, *contrast masking*), que depende también de la secuencia de referencia. Por último, las diferencias enmascaradas se pueden agrupar de diferentes maneras para ilustrar el error de percepción con diversas dimensiones (POOL), y el error agrupado se convierte en calidad visual (VQ, *visual quality*).

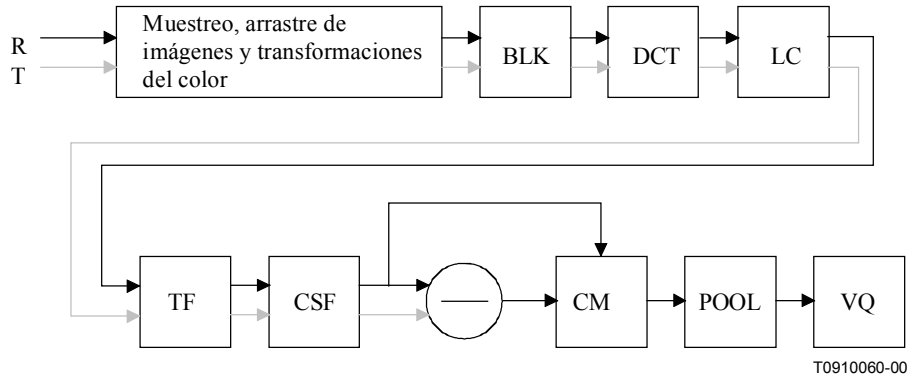


Figura VII.1/J.144 – Visión general de las etapas del procesamiento DVQ

Los parámetros de la medida se han estimado a partir de datos sicofísicos, tomados de publicaciones especializadas existentes y de mediciones de visibilidad de error de cuantificación de la DCT dinámica.

VII.2.1 Entrada

La entrada en la medición es un par de secuencias de imágenes en color. Sus dimensiones son {s,f,c,y,x}, donde s = secuencia (2), f = tramas, c = color (3), y = filas y x = columnas. La primera de las dos secuencias es la de referencia, la segunda es la de prueba. Normalmente, la de prueba diferirá de la de referencia por la presencia de los efectos visuales secundarios de la compresión. El espacio de color de entrada debe estar definido con suficiente detalle como para que pueda ser transformado en coordenadas de CIE, por ejemplo, especificando las coordenadas gamma y cromaticidad de cada primario. Dos ejemplos comunes utilizados en este apéndice son un espacio RGB lineal (gamma = 1) e YCbCr con gamma = 2,2.

VII.2.2 Transformaciones de color

El primer paso del proceso es la conversión de ambas secuencias de imágenes al espacio de color YOZ. Se trata de un espacio de color utilizado previamente al modelar errores de percepción en la compresión de imágenes fijas. Los tres componentes de este espacio son Y (luminancia de CIE

en candelas/m^2), O, un canal de color antagónico dado por $O=\{X=0{,}47; Y=-0{,}37; Z=-0{,}1\}$ y un canal azul dado por la coordenada Z de CIE. La transformación al espacio YOZ implica normalmente:

1) una transformación gamma, seguida por
2) una transformación de color lineal. Estas operaciones no alteran la dimensionalidad de la entrada.

VII.2.3 DCT por bloques

En este punto, se aplica una DCT por bloques a cada trama de cada canal de color. Las dimensiones del resultado son {s,f,c,by,bx,v,u}, donde by y bx son el número de bloques de las direcciones vertical y horizontal, siendo ahora v = u = 8.

VII.2.4 Contraste local

Los coeficientes de la DCT se convierten en unidades de contraste local de la manera siguiente. En primer lugar se extraen los coeficientes c.c. de todos los bloques. A continuación, se efectúa un filtrado temporal, utilizando un filtro IIR de paso bajo y primer orden con una ganancia de 1 y una constante de tiempo de τ_1. Los coeficientes DCT se dividen a continuación por los coeficientes DCT filtrados bloque por bloque. Los bloques Y y Z se dividen por los coeficientes c.c. de Y y Z; el bloque O se divide por el coeficiente c.c. de Y. En cada caso, se añade una constante muy pequeña al divisor para evitar la división por cero. Finalmente, los cocientes se ajustan mediante las magnitudes relativas de sus coeficientes correspondientes a una función básica de contraste unidad. Estas operaciones convierten cada coeficiente DCT en un número comprendido entre -1 y 1, que expresa la amplitud de la función básica correspondiente como una fracción de la luminancia media en ese bloque.

Los coeficientes c.c. se convierten de manera similar: sustrayendo el c.c. de valor medio de toda la trama, y dividiendo el resultado por ese valor medio.

VII.2.5 Filtrado temporal

Ambas secuencias se someten a continuación a un filtrado temporal. El filtro temporal es un filtro IIR de segundo orden, descrito más arriba al referirnos a los datos de ruido dinámico de la DCT. La utilización de un filtro IIR reduce al mínimo el número de tramas de datos que se han de retener en la memoria. Para mayor simplicidad, se puede utilizar un filtro de primer orden.

VII.2.6 Conversión de JND

Los coeficientes DCT, expresados ahora en forma de contraste local, se convierten a continuación en diferencias apenas perceptibles (JND) dividiéndolos por sus respectivos umbrales espaciales. Dichos umbrales se multiplican primero por un factor de suma espacial *s*, cuya finalidad y estimación se describe más adelante. Los umbrales de los dos canales de color se derivan de los umbrales de luminancia[3] o se basan en umbrales cromáticos adicionales. Tras convertirlos en JND, los coeficientes de las dos secuencias se sustraen unos de otros para generar una *secuencia de diferencias*.

VII.2.7 Enmascaramiento del contraste

El enmascaramiento del contraste se lleva a cabo construyendo primero una *secuencia de enmascaramiento*. Dicha secuencia empieza como la secuencia de referencia, después de la conversión de las JND. La secuencia es rectificada y a continuación sometida a un filtrado temporal mediante un filtro IIR discreto de paso bajo y primer orden, con una ganancia de g_1 y una constante de tiempo de τ_2. Estos valores se elevan seguidamente a la *m-ésima* potencia, cualquier valor inferior a 1 se sustituye por uno, y el resultado se utiliza para dividir la secuencia de diferencias. Con este proceso se reproduce de forma mímica el resultado del enmascaramiento del contraste tradicional en

el que los contrastes por debajo del umbral no tienen efecto enmascarador, y el efecto de los contrastes por encima del umbral aumenta con la *m-ésima* potencia del contraste de la máscara en diferencias apenas perceptibles (JND).

VII.2.8 Agrupamiento de Minkowski

Las dimensiones del resultado en este punto son {*f,c,by,bx,v,u*} donde, recordemos, *f* significa tramas, *c* significa canales de color, *by* y *bx* son el número de bloques en las direcciones vertical y horizontal y $v = u$ son las frecuencias vertical y horizontal. Estos errores elementales se pueden combinar a continuación tomando en consideración varias dimensiones, o todas ellas, para generar mediciones resumen del error visual. La suma correspondiente se efectúa utilizando una medida de Minkowski:

$$J_x = M\left(j_{f,c,by,bx,y,x},\beta\right) = \left(\sum_x \left|j_{f,c,by,bx,y,x}\right|^{\beta}\right)^{\frac{1}{\beta}} \qquad \text{(VII-1)}$$

En esta ecuación se ha indicado el sumatorio con las seis dimensiones, pero también se puede considerar cualquier subconjunto de las mismas. Una de las ventajas de la formulación de Minkowski es que se puede anidar. Por ejemplo, se suma primero teniendo en cuenta únicamente la dimensión de color *(c)*, y los resultados se suman a continuación teniendo en cuenta, por ejemplo, los números de bloques (*by* y *bx*).

VII.3 Evaluación

Se ha evaluado el comportamiento de la medida de la calidad vídeo DVQ comparando sus predicciones con los juicios respecto a la degradación formulados por 25 observadores que visionaron cinco secuencias de referencia procesadas por 12 circuitos ficticios de referencia (HRC). La medida DVQ es notablemente mejor que la de los modelos basados en la simple proporción de bits erróneos o en el error cuadrático medio (rms). La calidad de las predicciones hace pensar que la medida puede ser de utilidad en aplicaciones prácticas. Más recientemente presentamos nuestro algoritmo al Grupo de expertos en calidad vídeo (VQEG) que está elaborando un proyecto de prueba. La medida DVQ dio muy buenos resultados con una amplia gama de subconjuntos de HRC. Sobre todo en el régimen de calidad alta, con una correlación de rango de 0,72. Dos de las condiciones probadas, ciclos de registro/reproducción profesional de 1/2 pulgada multigeneración y errores de transmisión, quedan fuera del campo de aplicación de nuestro modelo. Suprimiendo esos HRC, la correlación de rango de Spearman fue de 0,82.

VII.4 Referencias

[1] WATSON (A.B.): Toward a perceptual video quality metric in Human Vision, *Visual Processing, and Digital Display VIII*, San Jose, CA: SPIE, Bellingham, WA, 1998.

[2] WATSON (A.B.), *et al*.: Design and performance of a digital video quality metric in Human Vision, *Visual Processing, and Digital Display IX.* San Jose, CA: SPIE, Bellingham, WA, 1999.

[3] WATSON (A.B.), HU (J.), McGOWAN (J.F.), III: *DVQ:* A digital video quality metric based on human vision, *Journal of Electronic Imaging*, 2000. En prensa.

APÉNDICE VIII

KPN/Swisscom CT

VIII.1 Introducción

En la medición de la percepción de la calidad vídeo (PVQM), las señales físicas de entrada y salida del dispositivo sometido a prueba (por ejemplo, un códec, o una cadena de transmisión) se convierten en representaciones psicofísicas (véase la figura VIII.1) que concuerdan lo más posible con las representaciones internas de las señales de audio/vídeo (representaciones dentro de nuestra cabeza). La calidad del dispositivo sometido a prueba se evalúa sobre la base de las diferencias en la representación interna. En PVQM, la representación interna, de la que se deriva la calidad, es tal que ambas distorsiones, la espacial y la temporal, quedan cubiertas por el método de medición.

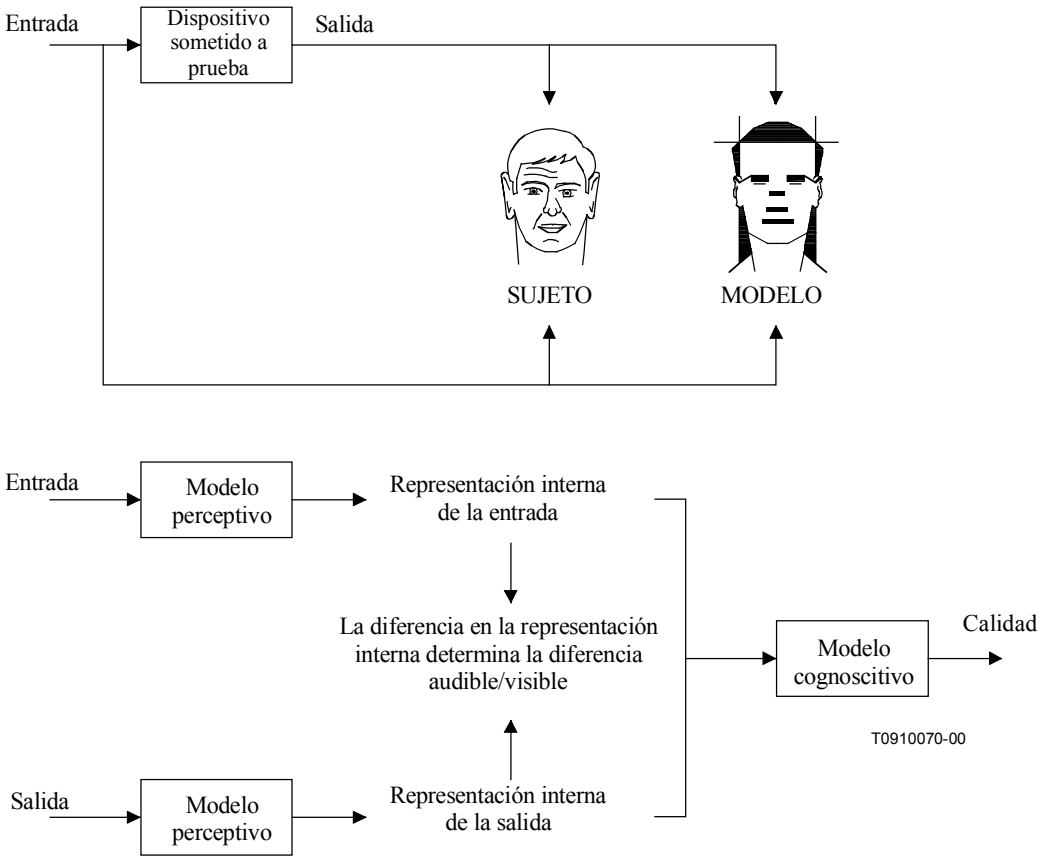


Figura VIII.1/J.144

Visión general de la filosofía básica aplicada en el desarrollo de PVQM. Se emplea un modelo del sujeto por computador, formado por un modelo perceptivo y un modelo cognoscitivo, para comparar la salida del dispositivo sometido a prueba (por ejemplo, un códec de vídeo) con la entrada, utilizando cualquier señal de vídeo.

Para hacer posible la utilización de PVQM en situaciones prácticas se incluye una alineación de la luminancia espacio-temporal en el algoritmo. Es bien sabido que los cambios globales del brillo y el contraste repercuten sólo de manera limitada en la calidad percibida subjetivamente, sobre todo cuando se compara con las consecuencias que tienen distorsiones tales como las derivadas de la división en bloques. Este efecto se cuantifica en PVQM utilizando una adaptación del brillo/contraste especial de la secuencia de vídeo distorsionada. Además, es evidente que sólo se puede efectuar un cálculo significativo de la medida de la distorsión si se sabe qué partes de las

señales de entrada y salida se han de comparar. Por ello, el algoritmo PVQM aplica un procedimiento de alineación espacio-temporal que busca la concordancia de los bloques antes de que se lleven a cabo las mediciones de manera efectiva.

La parte análisis de la luminancia espacial se basa en la detección del borde de la señal Y, mientras que la parte temporal se basa en el análisis de las tramas de diferencia de la señal Y. Es bien conocido el hecho de que el sistema visual humano (HVS) es mucho más sensible a la intensidad del componente luminancia que a la de los componentes crominancia. Además, el HVS tiene una función de sensibilidad al contraste que disminuye a frecuencias espaciales elevadas. Estas características básicas del HVS están reflejadas en el primer paso del algoritmo PVQM que proporciona una aproximación de primer orden a las funciones de sensibilidad al contraste de las señales de luminancia y crominancia.

En el segundo paso se calcula la intensidad del contorno de luminancia Y como una representación de la señal que contiene los aspectos más importantes de la imagen. La intensidad se computa calculando el gradiente local de la señal de luminancia en cada trama. El error relativo de intensidad del contorno entre el vídeo de entrada y el de salida se agrega a lo largo del espacio y el tiempo utilizando las medidas p de Lebesgue.

En el tercer paso, se calcula el error de crominancia como una media ponderada del error de color de ambos componentes Cb y Cr normalizados en la saturación local con predominio del componente Cr.

En el último paso, los tres indicadores diferentes se convierten en un indicador de calidad único, aplicando una sencilla una regresión lineal múltiple, que correlaciona adecuadamente la calidad vídeo global de la secuencia percibida de manera subjetiva. El método ha sido validado en KPN Research utilizando una gran variedad de bases de datos que contienen distorsiones de códec (MPEG, UIT-T H.263, etc.) y distorsiones generadas artificialmente. En todas esas bases de datos, la correlación entre los valores objetivos del algoritmo PVQM y las notas medias de opinión subjetivas es superior a 0,9.

VIII.2 Referencias

[1] BEERENDS (J.G.), HEKSTRA (A.P.): Objective measurement of video quality, *ITU-T Study Group 12, Document COM 12-7*, febrero de 1997.

APÉNDICE IX

NTIA

Introducción

El presente apéndice desarrolla plenamente el algoritmo utilizado para calcular una medida de la calidad vídeo (VQM, *video quality metric*) que se ajuste de cerca a los juicios de calidad subjetivos sobre las escenas de vídeo. Esta versión del algoritmo VQM contiene varias mejoras con respecto al modelo presentado al Grupo de expertos en calidad de vídeo (VQEG). Las mejoras se introdujeron antes de disponer de los datos subjetivos del VQEG [1]. Además de proporcionar estimaciones basadas en la percepción de la calidad subjetiva, independiente de la tecnología, el cálculo de la VQM no es muy complejo y se puede emplear en aplicaciones de supervisión de la calidad de servicio en tiempo real, de manera continua. Los resultados se presentan tras comparar la VQM con las notas medias de opinión de nueve pruebas subjetivas de doble estímulo distintas que abarcan gran número de escenas, sistemas de vídeo y tecnologías de codificación diferentes. Siete de esos conjuntos de datos contienen sobre todo escenas de vídeo procedentes de aplicaciones de radiodifusión de contribución con calidad y distribución con calidad (> 1,5 Mbit/s) mientras que dos

de los conjuntos de datos contienen sobre todo escenas de vídeo de aplicaciones multimedios (< 1,5 Mbit/s).

IX.1 Descripción del algoritmo VQM

El algoritmo VQM consiste en una combinación lineal de cuatro parámetros que han sido optimizados para la distancia de observación normalizada de seis veces la altura de la imagen. Tres de los parámetros se extraen de gradientes espaciales del componente luminancia (Y) de los trenes de vídeo de entrada y salida de UIT-R BT.601 [2], mientras que el tercer parámetro procede del vector formado por los componentes de crominancia (C_B, C_R).

Se supone que los trenes de vídeo de entrada y salida muestreados han sido calibrados antes de que se lleven a cabo los procesos que aquí se describen. La calibración incluye compensación de la ganancia del sistema y el desplazamiento del nivel, así como registro espacial y temporal de las imágenes.

IX.2 Parámetros de gradiente espacial

En la figura IX.1 se da una visión general del algoritmo utilizado para extraer parámetros de gradiente espacial. Los componentes Y de los trenes de vídeo de entrada y salida se procesan utilizando filtros de realce de bordes horizontales y verticales. Los trenes de vídeo procesados se dividen entonces en regiones espaciales-temporales (S-T) de las que se extraen las características, o las estadísticas resumen, que cuantifican la actividad espacial como una función de la orientación angular. A continuación, esas características se recortan en el extremo más bajo para emular umbrales de perceptibilidad. Seguidamente se calculan las distorsiones de la calidad de vídeo debidas a ganancias y pérdidas de los valores de las características para cada región S-T mediante la comparación de sus valores de entrada y salida utilizando relaciones funcionales que emulan el enmascaramiento visual de las degradaciones. Esas distorsiones se agrupan entonces a lo largo del espacio (colapso espacial) y el tiempo (colapso temporal) para producir los parámetros de calidad de una cuña de vídeo que, nominalmente, tienen una duración de 5 a 10 segundos.

Los filtros de realce de bordes, el tamaño de la región S-T y los umbrales de perceptibilidad que aquí se presentan se optimizaron en base a la correlación con las distorsiones de la percepción a una distancia de seis veces la altura de la imagen.

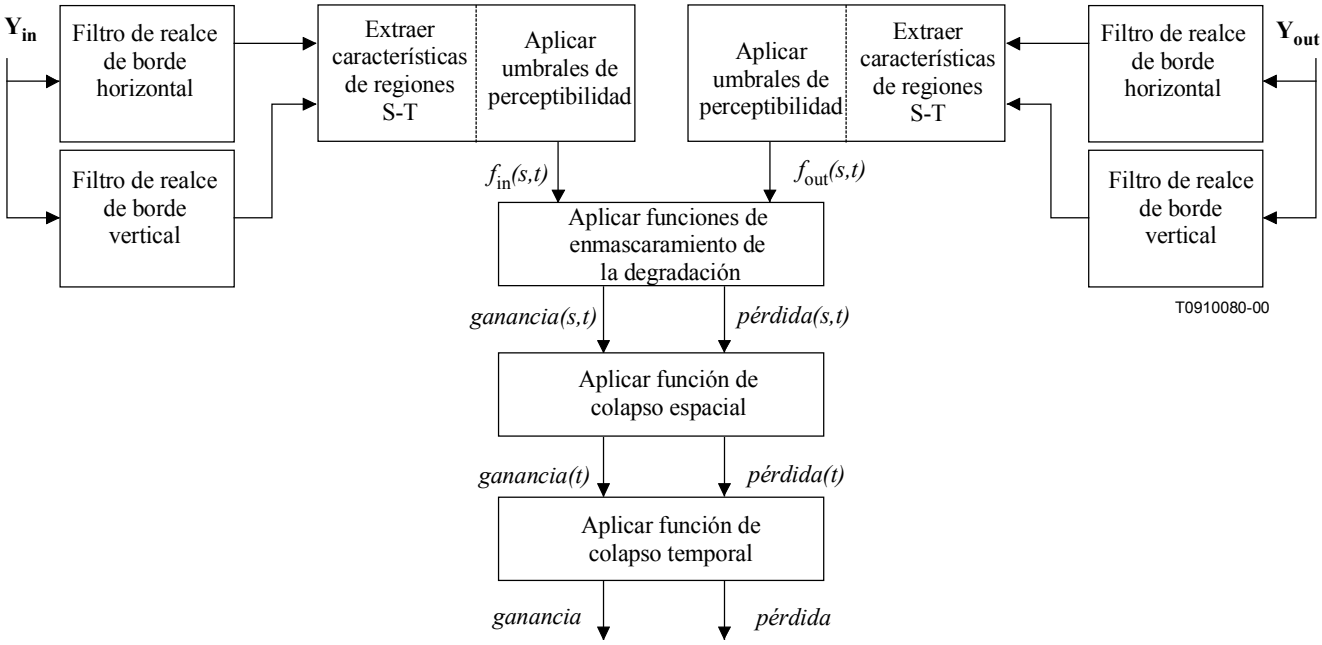


Figura IX.1/J.144 – Visión general del algoritmo utilizado para extraer los parámetros de gradiente espacial

IX.3 Filtros de realce de bordes

Las *tramas* de vídeo de entrada y salida se procesan primero con filtros de realce de los bordes horizontal y vertical que realzan los bordes al mismo tiempo que reducen el ruido. Los dos filtros mostrados en la figura IX.2 se aplican separadamente, uno para realzar diferencias de píxels horizontales mientras que se allanan las diferencias verticales (filtro de izquierda), y otro para realzar las diferencias de píxels verticales mientras que se allanan las diferencias horizontales (filtro de la derecha).

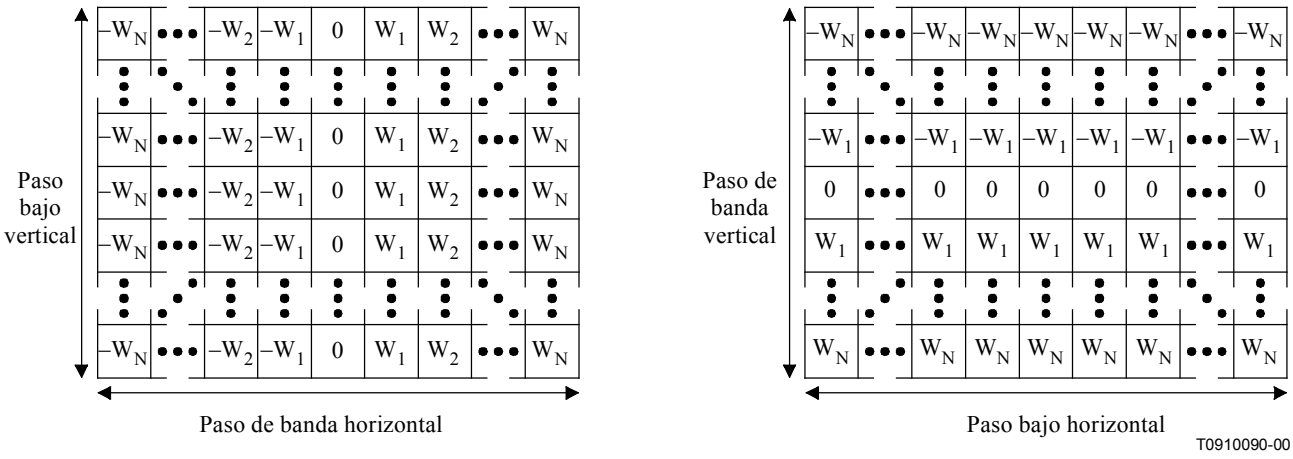


Figura IX.2/J.144 – Filtros de realce de bordes

Un filtro es algo así como la matriz transpuesta del otro; tienen un tamaño de 13 x 13 y ponderaciones de filtro dadas por

$$w_x = k * \left(\frac{x}{c}\right) * \exp\left\{-\frac{1}{2}\left(\frac{x}{c}\right)^2\right\}$$

donde x es el desplazamiento de un píxel con respecto al centro del filtro (0, 1, 2, ... , N), c es una constante que fija la anchura del filtro de paso de banda, y k es una constante de normalización elegida de tal manera que cada filtro produzca la misma ganancia que un filtro Sobel verdadero. Se ha encontrado que el grado óptimo de filtrado de paso de banda horizontal para una distancia de observación de seis veces la altura de la imagen viene dada por el filtro de $c = 2$, que tiene una respuesta de cresta de unos 4,5 ciclos/grado. Las ponderaciones de filtro de paso de banda utilizadas vienen dadas por:

[–0,0052625, –0,0173446, –0,0427401, –0,0768961, –0,0957739, –0,0696751, 0, 0,0696751, 0,0957739, 0,0768961, 0,0427401, 0,0173446, 0,0052625].

Obsérvese que los filtros de la figura IX.2 tienen una respuesta de paso bajo uniforme. Con esa respuesta se obtiene la mejor estimación de la calidad y tiene la ventaja añadida de su conveniencia desde el punto de vista del cálculo (por ejemplo, para el filtro de la izquierda de la figura IX.2, basta con sumar los píxels de una columna y multiplicar una vez por la ponderación).

IX.4 Tamaño de región S-T

Los trenes de vídeo de entrada y salida con bordes horizontal y vertical realzados se dividen, cada uno de ellos, en regiones S-T localizadas. La figura IX.3 da el tamaño de región S-T (8 píxels horizontales × 8 líneas verticales × 6 tramas de vídeo) con el que se alcanzó la correlación máxima con los índices subjetivos. Se señala, no obstante, que según se comprobó, esa correlación empeora *lentamente* a medida que uno se aleja del punto óptimo. Se pueden utilizar anchuras horizontales y verticales de hasta 32 píxels o líneas y anchuras temporales de hasta 30 tramas con resultados satisfactorios, lo que da al diseñador del sistema de medición objetiva un alto grado de flexibilidad para adaptar las técnicas aquí presentadas a los tamaños de regiones S-T diferentes.

Las características de cada región S-T se extraen calculando valores estadísticos resumidos a lo largo de la región S-T. En IX.5 se da una descripción detallada de las características extraídas.

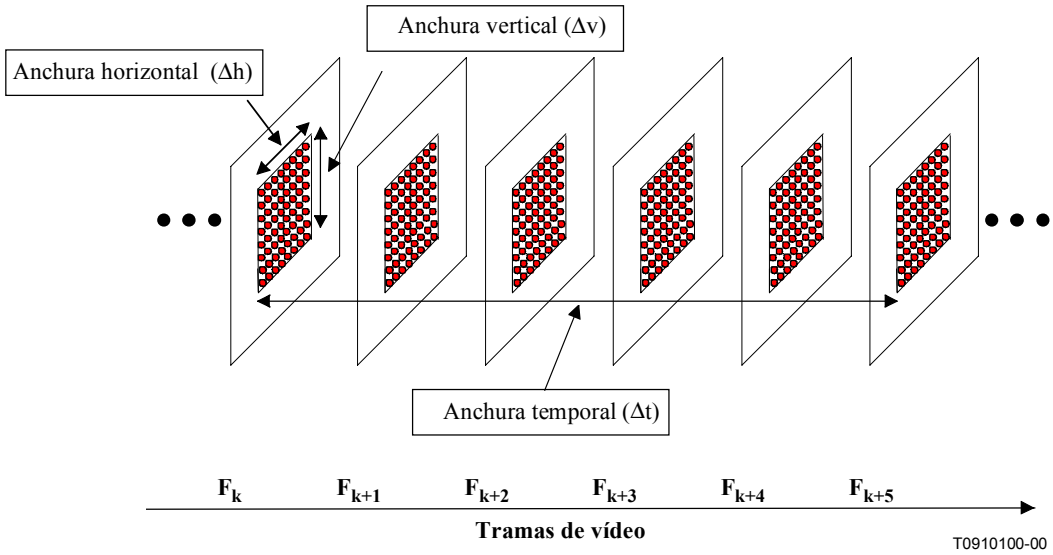


Figura IX.3/J.144 – Tamaño de región espacial-temporal óptima (S-T) para la extracción de características

IX.5 Descripción de las características

En esta cláusula se describe la extracción de dos características de actividad espacial de las regiones S-T de los trenes de vídeo de entrada y salida con bordes realzados a los que se refiere IX.4. El filtro mostrado en la figura IX.2 (izquierda) mejora los gradientes espaciales en la dirección horizontal (H) mientras que las transpuestas de ese filtro mejora los gradientes espaciales en la dirección vertical (V). La respuesta en cada píxel de los filtros H y V se puede representar en un diagrama bidimensional como el de la figura IX.4 en el que la respuesta del filtro H es el valor de abscisas y la respuesta del filtro V es el valor de ordenadas. Para un píxel de imagen dado situado en la fila *i*, columna *j*, y en el momento *t*, las respuestas de los filtros H y V se indicarán por $H(i, j, t)$ y $V(i, j, t)$, respectivamente. Estas respuestas se pueden convertir en coordenadas polares (R, θ) utilizando las relaciones:

$$R(i,j,t) = \sqrt{H(i,j,t)^2 + V(i,j,t)^2}$$

y:

$$\theta(i,j,t) = \tan^{-1}\left[\frac{V(i,j,t)}{H(i,j,t)}\right]$$

La primera característica, f_1, se calcula simplemente como desviación típica (*stdev*) a lo largo de la región S-T de las muestras $R(i, j, t)$ y se recorta a continuación al umbral de perceptibilidad P (es decir, si el resultado del cálculo *stdev* queda por debajo de P, f_1 se hace igual a P), esto es:

$$f_1 = \left\{ stdev\left[R(i,j,t)\right]\right\}\Big|_P : \; i,j,t \in \left\{\text{Región S-T}\right\}$$

Esta característica es sensible a los cambios en el volumen global de actividad espacial dentro de una región S-T dada. Por ejemplo, una borrosidad localizada produce una reducción del volumen de actividad espacial mientras que el ruido produce un aumento. El umbral P recomendado para esta característica es de 12.

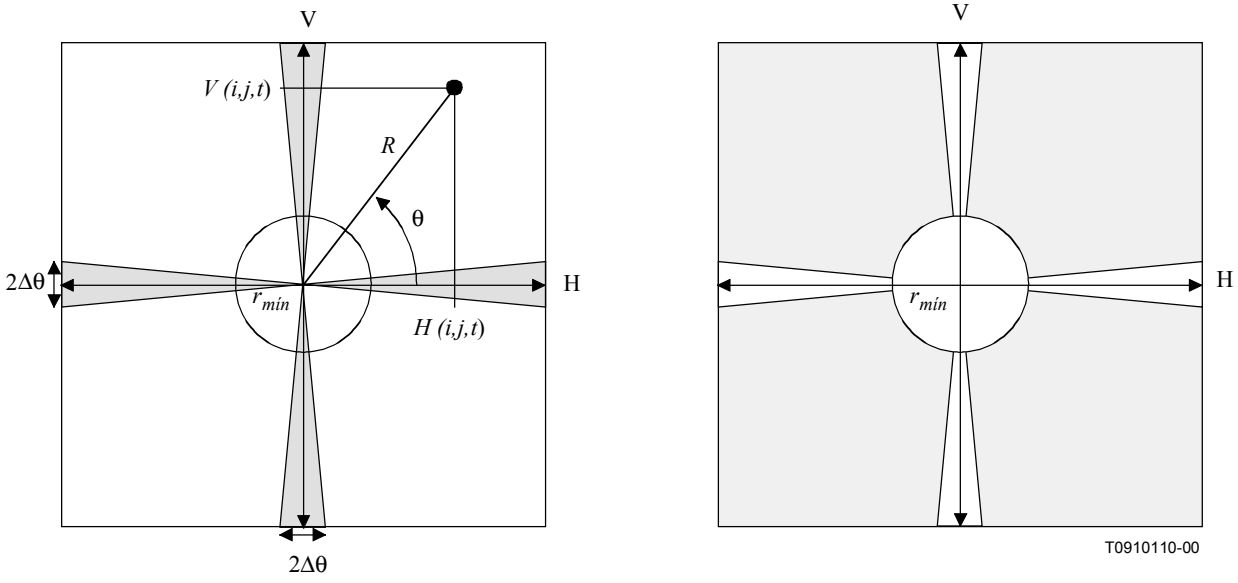


Figura IX.4/J.144 – División de la actividad espacial horizontal (H) y vertical (V) en distribuciones *HV* (izquierda) y $\overline{HV}$ (derecha)

La segunda característica, f_2, es sensible a los cambios en la distribución angular, u orientación, de la actividad espacial. Se calculan imágenes complementarias con las distribuciones de gradientes espaciales sombreados que se muestran en la figura IX.4. La imagen con gradientes horizontal y vertical, indicada por *HV*, contiene los píxels $R(i, j, t)$ que son bordes horizontales o verticales (a los píxels que son bordes diagonales se les da el valor cero). La imagen con los gradientes diagonales, indicada por $\overline{HV}$, contiene los píxels $R(i, j, t)$ que son bordes diagonales (a los píxels que son bordes horizontales o verticales se les da el valor cero). Las magnitudes de gradientes representados por $R(i, j, t)$ inferiores a $r_{mín}$ se igualan a cero en ambas imágenes para asegurar la exactitud de los cálculos de θ. Los píxels de *HV* y $\overline{HV}$ se pueden representar matemáticamente de la siguiente manera:

$$HV(i,j,t)=\begin{cases} R(i,j,t) \quad si\ R(i,j,t)\geq r_{mín}\ y\ m\frac{\pi}{2}-\Delta\theta<\theta(i,j,t)<m\frac{\pi}{2}+\Delta\theta \quad (m=0,1,2,3) \\ 0 \quad \text{de otro modo} \end{cases}$$

y:

$$\overline{HV}(i,j,t)=\begin{cases} R(i,j,t) \quad si\ R(i,j,t)\geq r_{mín}\ y\ m\frac{\pi}{2}+\Delta\theta\leq\theta(i,j,t)\leq(m+1)\frac{\pi}{2}-\Delta\theta \quad (m=0,1,2,3) \\ 0 \quad \text{de otro modo} \end{cases}$$

donde:

$$i,j,t\in\{\text{Región S-T}\}$$

Para el cálculo de los anteriores *HV* y $\overline{HV}$, el valor de $r_{mín}$ recomendado es 20 y el valor de Δθ recomendado es de 0,05236 radianes. La característica f_2 de una región S-T viene dada entonces por la relación entre el valor medio de *HV* y el valor medio de $\overline{HV}$, recortándose esos valores medios resultantes a sus umbrales de perceptibilidad *P*, a saber:

$$f_2=\{valor\ medio\ de[HV(i,j,t)]\}\big|_P \Big/ \{valor\ medio\ de[\overline{HV}(i,j,t)]\}\big|_P$$

El umbral de perceptibilidad recomendado *P* para el valor medio de *HV* y $\overline{HV}$ es de 3. La característica f_2 es sensible a los cambios en la distribución angular de la actividad espacial dentro de una región S-T dada. Por ejemplo, si los bordes horizontales y verticales experimentan más borrosidad que los bordes diagonales, f_2 de la salida será inferior a f_2 de la entrada. Por otro lado, si se introducen bordes horizontales o verticales erróneos, en forma, por ejemplo, de distorsiones debidas a la división en bloques o al enlosamiento, f_2 de la salida será superior a f_2 de la entrada. La característica f_2 proporciona así una manera sencilla de incluir las variaciones de la sensibilidad del sistema visual humano con respecto a la orientación angular.

En el análisis que sigue, el tren de características de entrada se indicará por $f_{in}(s, t)$ y el tren de características de salida correspondiente se indicará por $f_{out}(s, t)$, donde *s* y *t* son índices que denotan la posición espacial y la posición temporal, respectivamente, de la región S-T dentro de los trenes de vídeo de entrada y salida calibrados.

IX.6 Funciones de enmascaramiento de la degradación

A continuación se calcula la degradación de la percepción en cada región S-T utilizando una función que modela el enmascaramiento visual de las degradaciones. La ganancia y la pérdida se deben examinar por separado, ya que producen básicamente efectos diferentes en la percepción de la calidad (por ejemplo, pérdida de la actividad espacial debido a la borrosidad y ganancia de la actividad espacial debida al ruido o la división en bloques). Dos de las muchas funciones de comparación que se han evaluado, han producido consistentemente la mejor correlación con los índices subjetivos. Esas funciones de comparación modelan la perceptibilidad de las degradaciones espaciales o temporales. Para una región S-T dada, las distorsiones de la ganancia o la pérdida se calculan utilizando:

$$ganancia(s,t) = pp\left\{\log_{10}\left[\frac{f_{out}(s,t)}{f_{in}(s,t)}\right]\right\}$$

y:

$$pérdida(s,t) = np\left\{\frac{f_{out}(s,t) - f_{in}(s,t)}{f_{in}(s,t)}\right\}$$

donde *pp* es el operador de la parte positiva (es decir, los valores negativos son sustituidos por cero), y *np* es el operador de la parte negativa (es decir, los valores positivos son sustituidos por cero). Estas funciones de enmascaramiento visual implican que la percepción de la degradación es inversamente proporcional al volumen de actividad espacial o temporal localizada en la escena de entrada. En otras palabras, las degradaciones espaciales se hacen menos visibles a medida que aumenta la actividad espacial en la escena de entrada (es decir, a medida que aumenta el enmascaramiento espacial), y las degradaciones temporales se hacen menos visibles a medida que aumenta la actividad temporal en la escena de entrada (es decir, a medida que aumenta el enmascaramiento temporal). Si bien las funciones de comparación logarítmica y proporcional se comportan de manera muy similar, la función logarítmica tiende a ser ligeramente más ventajosa para las ganancias mientras que la función proporcional tiende a ser ligeramente más ventajosa para las pérdidas.

IX.7 Función de colapso espacial

A continuación se agrupan las degradaciones de las regiones S-T con el mismo índice de tiempo *t* utilizando una función de colapso espacial. Profundas investigaciones han puesto de manifiesto que las funciones de colapso espacial óptimas conllevan normalmente alguna forma de procesamiento del caso más degradado. Esto se debe a que las degradaciones localizadas tienden a llamar la atención del observador, haciendo que la parte más degradada de la imagen sea el factor predominante en la decisión sobre calidad subjetiva. La función de colapso espacial se calcula para cada índice temporal *t* como el valor medio del peor 5% de las distorsiones medidas con el índice espacial *s* (indicado como $peor\_5\%_{espacio}$). Esto equivale a clasificar las distorsiones de ganancia para cada índice temporal *t* y promediar las distorsiones que están por encima del umbral del 95%. De manera similar, las distorsiones de pérdida se clasifican para cada índice temporal *t*, pero se utiliza el valor medio de las distorsiones que están por debajo del umbral del 5% (porque las pérdidas son negativas). Aplicando la función $peor\_5\%_{espacio}$ se obtiene el historial de las muestras de ganancia y pérdida, a saber *ganancia*(*t*) y *pérdida*(*t*), que a continuación debe ser colapsado temporalmente.

IX.8 Función de colapso temporal

Por último, los resultados de la función de colapso espacial se agrupan utilizando una función de colapso temporal para generar un parámetro objetivo de la cuña de vídeo, cuya duración nominal es de 5 a 10 segundos. Los observadores parecen utilizar varias funciones de colapso temporal cuando evalúan subjetivamente cuñas de vídeo cuya duración es de 9 a 10 segundos. Una función de colapso temporal indica el nivel de calidad medio de la cuña mientras que la otra indica la peor calidad transitoria de la cuña (por ejemplo, los errores de transmisión digital provocan normalmente una perturbación de 1 a 2 segundos en el vídeo de salida).

El valor medio a lo largo del tiempo (indicado por *medio* $_{\text{tiempo}}$) parece ser indicativo de la calidad media observada durante el periodo de tiempo. Para la calidad transitoria peor, el nivel del 10% a lo largo del tiempo de los parámetros de pérdida (indicado por $10\%_{\text{tiempo}}$) y el nivel del 90% a lo largo del tiempo de los parámetros de ganancia (indicado por $90\%_{\text{tiempo}}$) parecen reflejar de la mejor manera posible la repercusión subjetiva (se clasifican las muestras del historial del parámetro de pérdidas y se utiliza el nivel del 10%; se clasifican las muestras del historial del parámetro de ganancias y se utiliza el nivel del 90%). Es preciso seguir investigando para optimizar estas funciones de colapso temporal.

IX.9 Tres parámetros de gradiente espacial

Los tres parámetros de gradiente espacial utilizados para computar el algoritmo VQM vienen dados por:

f_1*_pérdida* (utilizar la función de colapso temporal $10\%_{\text{tiempo}}$);

f_2*_pérdida* (utilizar la función de colapso temporal *medio*$_{\text{tiempo}}$), y

f_2*_ganancia* (utilizar la función de colapso temporal *medio*$_{\text{tiempo}}$).

Las características f_1 y f_2 se describen en IX.5, las funciones de pérdida de ganancia se dan en IX.6, la función de colapso espacial se da en IX.7 y la función de colapso temporal figura en IX.8.

IX.10 Parámetros de crominancia

Esta cláusula presenta un único parámetro de distorsión de la crominancia que se incluye en el cómputo de VQM. Para un píxel de imagen dado, situado en la fila i, columna j, y en el momento t, sean $C_B(i, j, t)$ y $C_R(i, j, t)$ valores de C_B y C_R de UIT-R BT.601. Los componentes de un vector de la característica de crominancia bidimensional, $\underline{f}_C$, se calculan simplemente como el valor medio (*medio*) a lo largo de la región S-T de las muestras $C_B(i, j, t)$ y $C_R(i, j, t)$ respectivamente, dando una mayor ponderación perceptiva al componente C_R:

$$\underline{f}_C(s,t) = (valor\ medio\ de[C_B(i,j,t)], W_R * valor\ medio\ de[C_R(i,j,t)]) :\ i,j,t \in \{\text{Región S-T}\},\ \text{y}$$

$$W_R = 1{,}5\,.$$

El tamaño recomendado de la región S-T es de 8 píxels en la dirección horizontal × 8 líneas en la dirección vertical × 1 trama de vídeo (de hecho, 4 píxels de C_B y C_R en la dirección horizontal, ya que estas señales son submuestreadas por dos en UIT-R BT.601). La distorsión de crominancia de cada región S-T, indicada por $d_C(s, t)$, donde s y t son índices que denotan las posiciones espacial y temporal, respectivamente, de la región S-T dentro de los trenes de vídeo de entrada y salida calibrados, se calcula como la distancia euclideana entre los vectores de la característica de crominancia de entrada y salida $\underline{f}_{\text{Cin}}$ y $\underline{f}_{\text{Cout}}$ señalada por la línea de puntos de la figura IX.5, a saber:

$$d_C(s,t) = \left\| \underline{f}_{Cout}(s,t) - \underline{f}_{Cin}(s,t) \right\|$$

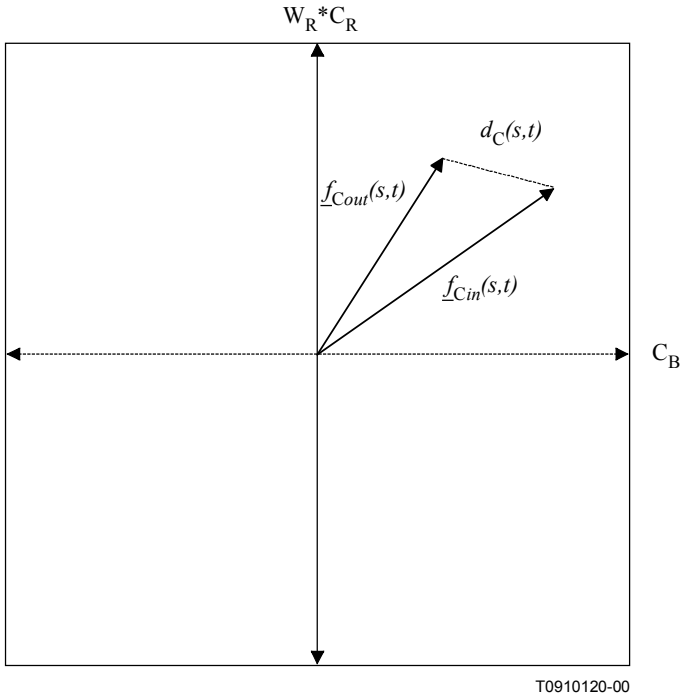


Figura IX.5/J.144 – Cálculo de la distorsión de crominancia d_C (*s, t*) de una región S-T

La función de colapso espacial óptima para d_C (*s*, *t*) es la desviación típica en el espacio (indicada por $stdev_{espacio}$), que es similar a la función $peor\_5\%_{espacio}$ indicada previamente. La función de colapso temporal óptima es el nivel de 10% a lo largo del tiempo (indicada por $10\%_{tiempo}$), que representa el nivel de la distorsión que está presente casi siempre. El valor de distorsión de la crominancia después del colapso espacial y temporal se recorta hasta un umbral de perceptibilidad de $P = 0{,}8$ y a continuación este valor recortado se sustrae para generar la medida d_C. Resumiendo, el parámetro de distorsión de crominancia d_C viene dado por:

$$d_C = \left\{10\%_{\text{tiempo}}\left[stdev_{\text{espacio}}\left(d_C(s,t)\right)\right]\right\}\Big|_P - P$$

IX.11 Cálculo de VQM

VQM se calcula aplicando la siguiente fórmula:

$$\text{VQM} = -0{,}3609*f_1\_pérdida + 0{,}5031*(f_2\_pérdida)^2 + 0{,}1390*f_2\_ganancia + 0{,}0295*d_C$$

La elevación a la segunda potencia del parámetro $f_2\_pérdida$ se necesita para linealizar esta respuesta de parámetro. El parámetro $f_1\_pérdida$ requiere un multiplicador negativo ya que siempre es inferior o igual a cero (el parámetro $f_2\_pérdida$ siempre es inferior o igual a cero pero en el cálculo de VQM se utiliza el cuadrado de este parámetro). Los parámetros $f_2\_ganancia$ y d_C son siempre superiores o iguales a cero. Calculada la medida de la calidad de vídeo (VQM) de esta manera, tendrá valores superiores o iguales a cero y su valor nominal máximo es de uno. Ocasionalmente puede ser superior a uno en escenas de vídeo sumamente distorsionadas.

IX.12 Descripción de conjuntos de datos subjetivos

Los nueve experimentos subjetivos se llevaron a cabo de 1992 a 1999. Los conjuntos de datos se obtuvieron de acuerdo con la versión más reciente de UIT-R BT.500-9 [3] de la que se disponía cuando se realizaron las pruebas. En todos los conjuntos de datos se utilizaron escenas de 9 a

10 segundos de duración y se aplicó observación de doble estímulo (los observadores vieron la secuencia original y la secuencia degradada). Para abreviar, sólo se da aquí un resumen de cada experimento subjetivo. Se remite al lector a las referencias de acompañamiento en donde podrá encontrar una descripción más completa.

Conjunto de datos uno [4, 5]

Un grupo de 48 observadores evaluó un total de 132 cuñas de vídeo que se generaron mediante emparejamiento aleatorio y determinístico de 36 escenas de prueba con 27 sistemas de vídeo. Las 36 escenas de prueba contenían cantidades de información espacial y temporal muy variables. Los 27 sistemas de vídeo incluían sistemas de compresión de vídeo digital funcionando a velocidades binarias de 56 kbit/s a 45 Mbit/s con tasas de error controladas, ciclos de codificación/decodificación NTSC, ciclos de grabación/reproducción VHS y S-VHS y transmisión por ondas métricas. A los observadores se les mostró primero la versión original y después la versión degradada y se les pidió que evaluaran la diferencia de calidad percibida utilizando la escala de cinco notas de degradación (imperceptible, perceptible pero no molesta, ligeramente molesta, molesta y muy molesta).

Conjunto de datos dos [6, 7]

Los grupos de observadores, formados por un total de 30 observadores de tres laboratorios diferentes, evaluaron 600 cuñas generadas emparejando 25 escenas de prueba con 24 sistemas de vídeo. Las 25 escenas de prueba incluían escenas de cinco categorías:

1) una persona, sobre todo su cabeza y sus hombros;
2) una persona con gráficos y/o más detalles;
3) más de una persona;
4) gráficos con un puntero; y
5) notable movimiento del objeto y/o la cámara.

Los 24 sistemas de vídeo incluían sistemas de teleconferencia por vídeo patentados y normalizados, funcionando a velocidades binarias de 56 kbit/s a 1,5 Mbit/s con tasas de error controladas, un códec a 45 Mbit/s y ciclo de grabación/reproducción VHS. El procedimiento de prueba subjetiva fue el mismo que el del conjunto de datos uno.

Conjunto de datos tres [8]

Un grupo de 32 observadores evaluó la diferencia de calidad entre escenas de entrada con cantidades de ruido añadido controladas y la salida procesada con el sistema de compresión MPEG 2. El conjunto de datos contiene un total de 105 cuñas de vídeo que fueron generadas emparejando siete escenas de prueba a tres niveles de ruido diferentes con cinco sistemas de vídeo MPEG-2. Las siete escenas de prueba se eligieron de modo que se abarcara toda una gama de movimientos, brillo, contraste y detalles espaciales. Los cinco sistemas de vídeo MPEG-2 funcionaron a velocidades binarias de 1,8 Mbit/s a 13,9 Mbit/s. A los observadores se les mostró la entrada y la salida procesada en un orden A/B aleatorizado y se les pidió que evaluaran la calidad de B utilizando A como referencia. En el experimento se utilizó una escala de comparación de siete puntos (B mucho peor que A, B peor que A, B ligeramente peor que A, B igual que A, B ligeramente mejor que A, B mejor que A, B mucho mejor que A).

Conjunto de datos cuatro [9]

Un grupo de 32 observadores evaluó un total de 112 cuñas de vídeo que se generaron emparejando subgrupos de ocho escenas (el número total de escenas en la prueba fue de 16), cada una de ellas con 14 sistemas de vídeo diferentes. Las 16 escenas de prueba abarcaban una amplia gama de movimientos, brillo, contraste y detalles espaciales e incluían material de películas, deportes, la naturaleza y escenas de prueba clásicas de UIT-R BT.601. Los 14 sistemas de vídeo incluían sistemas MPEG-2 que funcionaban a velocidades binarias de 2 Mbit/s a 36 Mbit/s con tasas de error

controladas, MPEG-2 de multigeneración, ciclos de grabación/reproducción profesional de 1/2 pulgada de multigeneración, VHS y sistemas de teleconferencia de vídeo funcionando a velocidades binarias de 768 kbit/s a 1,5 Mbit/s. El procedimiento de prueba subjetiva fue el mismo que el del conjunto de datos tres.

Conjunto de datos cinco [9]

Un grupo de 32 observadores evaluó un total de 42 cuñas de vídeo que se generaron emparejando subgrupos de seis escenas (el número total de escenas de la prueba fue de 12), cada una con siete sistemas MPEG-2 diferentes. Las 12 escenas de prueba incluían material de deportes y escenas de prueba clásicas de UIT-R BT.601. Los nueve sistemas MPEG-2 funcionaron a velocidades binarias de 2 Mbit/s a 8 Mbit/s. El procedimiento de pruebas subjetiva fue el mismo que el del conjunto de datos tres.

Conjunto de datos seis a nueve [10]

Cuatro conjuntos de datos (525 líneas y alta calidad, 525 líneas y baja calidad, 625 líneas y alta calidad, 625 líneas y baja calidad), cada uno de los cuales constaba de 90 cuñas de vídeo que se generaron emparejando diez escenas con nueve sistemas de vídeo. Para cada conjunto de datos, un total de 60 a 80 observadores de cuatro laboratorios diferentes (esto es, 15 a 20 observadores por laboratorio) evaluaron la calidad subjetiva utilizando la escala de calidad continua de doble estímulo (DSCQS). Las veinte escenas de prueba diferentes (diez para 525 líneas y 10 para 625 líneas) incluían material de deportes, escenas de prueba clásicas de UIT-R BT.601, gráficos en movimiento e imágenes fijas. Los sistemas de vídeo incluían sistemas MPEG-2 funcionando a velocidades binarias de 2 Mbit/s a 50 Mbit/s, sistemas de teleconferencia de vídeo funcionando a 768 kbit/s y 1,5 Mbit/s, algunos sistemas con errores de transmisión digital, MPEG-2 de multigeneración y ciclos de grabación/reproducción profesional de 1/2 pulgada de multigeneración en donde se utilizaron formatos de señal compuesta y/o de componentes.

IX.13 Resultados

En el cuadro IX.1 se da el coeficiente de correlación lineal de Pearson entre VQM y cada uno de los conjuntos de datos subjetivos. VQM alcanzó un coeficiente de correlación de Pearson medio de 0,90.

La figura IX.6 muestra el trazado disperso de los juicios de calidad subjetivos derivados de los nueve conjuntos de datos frente a VQM. En ese trazado, se ha hecho que las notas medias de opinión subjetiva de los nueve conjuntos de datos queden comprendidas entre cero y uno. El coeficiente de correlación lineal de Pearson entre las notas subjetivas y VQM del trazado disperso es de 0,94 (coeficiente de correlación superior a la media de los valores del cuadro IX.1 ya que la gama de calidad del conjunto de datos combinados es superior a la de cualquiera de los conjuntos de datos individuales). La mayoría de los puntos externos del trazado disperso proceden de sistemas en cuya salida hay alguna forma de ruido variable con el tiempo (por ejemplo, transmisión en ondas métricas, ciclos de grabación/reproducción profesional de 1/2 pulgada de multigeneración, ciclos de codificación compuestos, errores de transmisión digital que producen bloques de errores transitorios). Se está preparando la introducción de mejoras en el algoritmo VQM cara al futuro, entre ellas, parámetros basados en la percepción para medir los efectos del ruido variable con el tiempo. Los parámetros de calidad derivados de la información de gradiente temporal (es decir, la actividad temporal) constituyen un campo de investigación prometedor.

Cuadro IX.1/J.144 – Coeficiente de correlación lineal de Pearson para VQM

Conjunto de datos	Coeficiente de correlación lineal de Pearson
Uno	0,92
Dos	0,90
Tres	0,94
Cuatro	0,88
Cinco	0,91
Seis a nueve combinados	0,86

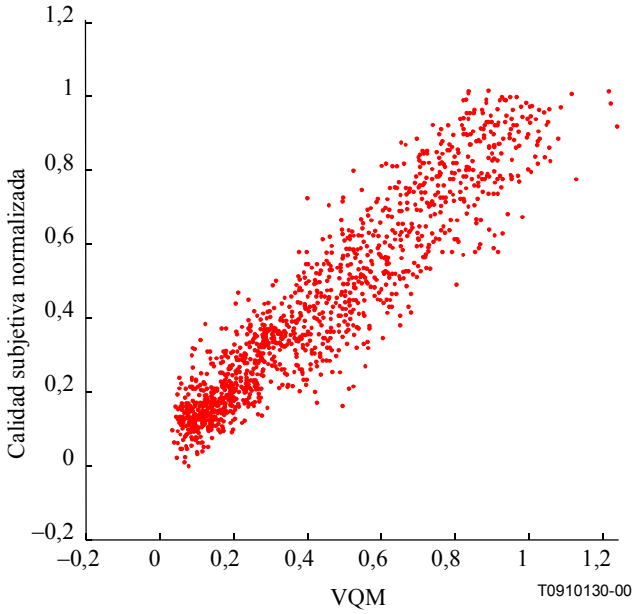


Figura IX.6/J.144 – Trazado disperso de los juicios de calidad subjetivos frente a VQM para nueve conjuntos de datos

IX.14 Referencias

[1] WOLF (Stephen), PINSON (Margaret H.): Spatial-temporal distortion metrics for in-service quality monitoring on any digital video system, *SPIE International Symposium on Voice, Video, and Data Communications*, Boston, MA, 11-22 de septiembre de 1999.

[2] UIT-R BT.601-5 (1995), *Parámetros de codificación de televisión digital para estudios con formato de imagen normal 4:3 de pantalla ancha 16:9*.

[3] UIT-R BT.500-9 (1998), *Metodología para la evaluación subjetiva de la calidad de las imágenes de televisión*.

[4] VORAN (Stephen), WOLF (Stephen): The Development and evaluation of an objective video quality assessment system that emulates human viewing panels, *International Broadcasting Convention (IBC)*, julio de 1992.

[5] WEBSTER (Arthur A.), JONES (Coleen T.), PINSON (Margaret H.), VORAN (Stephen D.), WOLF (Stephen): An objective video quality assessment system based on human perception, *Human Vision, Visual Processing, and Digital Display IV, Proceedings of the SPIE*, Vol. 1913, febrero de 1993.

[6] ANSI Accredited Standards Working Group T1A1 contribution number T1A1.5/94-118R1, "*Subjective test plan (tenth and final draft)*", Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington, DC, 3 de octubre de 1993.

[7] ANSI T1.801.01 (1995), *Digital Transport of Video Teleconferencing/Video Telephony Signals – Video Test Scenes for Subjective and Objective Performance Assessment*.

[8] FENIMORE (Charles), *et al.*: Perceptual effects of noise in digital video compression, *SMPTE Journal*, Vol. 109, pp. 178-186, marzo de 2000.

[9] WOLF (S.), PINSON (M.): In-service performance metrics for MPEG-2 video systems, *Made to Measure 98 – Measurement Techniques of the Digital Age Technical Seminar, technical conference jointly sponsored by the International Academy of Broadcasting (IAB), ITU, and the Technical University of Braunschweig (TUB)*, Montreux, Suiza, 12-13 de noviembre de 1998.

[10] Final report from the video quality experts group (VQEG) on the validation of objective models of video quality assessment, *VQEG meeting number 4*, Ottawa, Canadá, marzo de 2000.

SERIES DE RECOMENDACIONES DEL UIT-T

Serie A	Organización del trabajo del UIT-T
Serie B	Medios de expresión: definiciones, símbolos, clasificación
Serie C	Estadísticas generales de telecomunicaciones
Serie D	Principios generales de tarificación
Serie E	Explotación general de la red, servicio telefónico, explotación del servicio y factores humanos
Serie F	Servicios de telecomunicación no telefónicos
Serie G	Sistemas y medios de transmisión, sistemas y redes digitales
Serie H	Sistemas audiovisuales y multimedios
Serie I	Red digital de servicios integrados
Serie J	**Redes de cable y transmisión de programas radiofónicos y televisivos, y de otras señales multimedios**
Serie K	Protección contra las interferencias
Serie L	Construcción, instalación y protección de los cables y otros elementos de planta exterior
Serie M	RGT y mantenimiento de redes: sistemas de transmisión, circuitos telefónicos, telegrafía, facsímil y circuitos arrendados internacionales
Serie N	Mantenimiento: circuitos internacionales para transmisiones radiofónicas y de televisión
Serie O	Especificaciones de los aparatos de medida
Serie P	Calidad de transmisión telefónica, instalaciones telefónicas y redes locales
Serie Q	Conmutación y señalización
Serie R	Transmisión telegráfica
Serie S	Equipos terminales para servicios de telegrafía
Serie T	Terminales para servicios de telemática
Serie U	Conmutación telegráfica
Serie V	Comunicación de datos por la red telefónica
Serie X	Redes de datos y comunicación entre sistemas abiertos
Serie Y	Infraestructura mundial de la información y aspectos del protocolo Internet
Serie Z	Lenguajes y aspectos generales de soporte lógico para sistemas de telecomunicación

Impreso en Suiza
Ginebra, 2002