INFORME 1206

MÉTODOS DE EVALUACIÓN DE LA CALIDAD DE LA IMAGEN EN RELACIÓN CON LAS DEGRADACIONES DEBIDAS A LA CODIFICACIÓN DIGITAL DE LAS SEÑALES DE TELEVISIÓN

(Cuestión 3/11 y Programa de Estudios 3B/11)

(1990)

1. Introducción

La evaluación de las degradaciones debidas a la codificación tiene una importancia crítica debido a la aplicación creciente de la codificación digital y la transmisión a velocidad binaria reducida. La comprensión de los métodos de evaluación es importante no sólo desde el punto de vista de la calidad de funcionamiento de nuevos equipos de codificación, sino también del de la interpretación de las mediciones hechas sobre tales equipos y el de la especificación de objetivos de calidad. Por otra parte, los códecs digitales, al igual que todos los procesos digitales autoadaptables o no lineales, no pueden caracterizarse totalmente con las cartas o señales de prueba de televisión tradicionales.

De los estudios efectuados en relación con la Cuestión 3/11 y el Programa de Estudios 3B/11 se deduce que es deseable establecer relaciones entre las mediciones objetivas de señales degradadas por la codificación digital y las evaluaciones subjetivas de la calidad de la imagen así obtenida. En este Informe se avanza hacia ese objetivo, que se revela más difícil de lograr a medida que aumenta la complejidad de los códecs.

En la Recomendación 500 ———— se indican métodos subjetivos para evaluar la calidad de imagen en la televisión de resolución convencional así como su degradación. Para TVAD se indican en la Recomendación 710. En el presente Informe se analiza la aplicación de esos métodos a la evaluación de los códecs de televisión.

Ultimamente se ha adquirido una gran experiencia en lo relativo a la evaluación de la calidad de funcionamiento de códecs de alta calidad para televisión de componentes con relación 4:2:2 a 34, 45 y 140 Mbit/s [CCIR, 1986-90a]. En las correspondientes pruebas, se examinó la calidad de funcionamiento de los códecs en términos de calidad de imagen decodificada básica, calidad después del tratamiento posterior en estudio (incrustación cromática y cámara lenta) aplicado a las imágenes decodificadas y la degradación de la imagen decodificada, asociada con la presencia de una gama de proporciones de errores de canal. Algunas partes de este Informe se benefician de esas pruebas.

Las especificaciones de calidad en el caso de aplicaciones de distribución pueden expresarse en términos de la apreciación subjetiva de los observadores. En teoría, por tanto, esos códecs pueden ser evaluados subjetiva u objetivamente, contrastándolos con estas especificaciones. Sin embargo, la calidad de un códec diseñado para aplicaciones de contribución podría especificarse teóricamente en términos de parámetros objetivos de calidad de funcionamiento, porque su salida no está destinada a una visualización inmediata sino a tratamiento posterior en estudio, almacenamiento y/o codificación para transmisión ulterior. Dada la dificultad de definir esa calidad de funcionamiento para una diversidad de operaciones de tratamiento posterior, el enfoque preferido ha sido especificar la calidad de funcionamiento de una cadena de equipo, incluyendo una función de tratamiento posterior, a la que se considera representativa de una aplicación práctica de contribución. Esta cadena podría constar típicamente de un códec, seguido por una función de tratamiento posterior de estudio (o de otro códec en el caso de evaluación de calidad de contribución básica) seguido todavía por otro códec antes de que la señal alcance al observador. La adopción de esta estrategia para las especificaciones de códecs destinados a aplicaciones de contribución significa que los procedimientos de medición que se dan en el presente Informe pueden tambien utilizarse para su evaluación.

A lo largo del Informe se insiste en la importancia de elegir secuencias de imágenes de prueba críticas, sobre todo de escenas naturales, y se dan algunas directrices sobre cómo generar o escoger tales secuencias.

2. Clasificación de los códecs digitales

La función de la codificación digital es reducir la velocidad binaria necesaria para representar una secuencia de imágenes, asegurando al mismo tiempo una mínima pérdida de la calidad de imagen. El equipo de codificación hace esto eliminando, en primer lugar, tanta redundancia estadística de las imágenes como se pueda (es decir, que no se pierde calidad como consecuencia de esta primera etapa conceptual). A continuación, si hace falta más reducción de velocidad binaria, hay que introducir un cierto grado de distorsión en la imagen, si bien uno de los objetivos de diseño de los códecs es ocultar esa distorsión, aprovechando determinadas insensibilidades de percepción del sistema de visión humano.

Conviene dividir los códecs en dos clases, los que utilizan **codificación de longitud de palabra fija** y los que utilizan **codificación de longitud de palabra variable** (véanse las definiciones en los puntos 3.1. y 3.2 respectivamente). La segunda clase es más eficaz y compleja e incluye todos los sistemas propuestos recientemente para codificar vídeo con relación 4:2:2 en la gama 30-45 Mbit/s. La primera clase es no obstante suficiente para permitir la reducción de la señal de vídeo con relación 4:2:2 a 140 Mbit/s, preservando al mismo tiempo la calidad exigida para aplicaciones de contribución. Resulta útil subdividir también estas clases de códecs en códecs **intratrama** (o espaciales) y códecs **intercuadro** (incluidos los intertrama), que contienen almacenamientos de cuadros (o tramas) que les permite aprovechar la redundancia que existe entre cuadros (o tramas) de imágenes sucesivas.

Está apareciendo una tercera clase de códec que emplea la codificación de longitud de palabra variable pero que está diseñado para redes de velocidad binaria variable. Estos códecs pueden en principio mantener una calidad de imagen decodificada constante, sujeta a los límites de los valores máximos de la demanda máxima de la red. Las pruebas de calidad de tales códecs, que están pendiente de estudio, deberá tener en cuenta la naturaleza de la red utilizada y las estadísticas de los datos inyectados por todos sus usuarios.

3. Evaluaciones objetivas de los códecs en términos de degradaciones de imagen percibidas

3.1 Códecs de longitud de palabra fija

Con los códecs de longitud de palabra fija se utiliza un número fijo de bits para representar un número fijo de muestras de imagen de origen. Por ejemplo, en los códecs de MIC o MICD con longitud de palabra fija, se atribuye un número fijo de bits a cada muestra de imagen y en los códecs de cuantificación vectorial o de transformación con longitud de palabra fija se atribuye a cada bloque de muestras de imagen un número fijo de bits.

3.1.1 Métodos basados en la utilización de señales de prueba sintéticas

En estos códecs, la degradación introducida en cada muestra de imagen recibida de una imagen, depende de los valores de las muestras ubicadas en su entorno, ya sea en la misma trama (para un códec intratrama) o en la misma trama y tramas anteriores (para un códec intercuadro). Es posible por tanto, utilizando señales de prueba digitales bidimensionales o tridimensionales adecuadamente elegidas, provocar de manera artificial las degradaciones características de la codificación digital de imágenes.

A algunos de estos **factores de degradación** se les han adjudicado nombres relacionados con su interpretación por los observadores tales, como, formación incorrecta de contornos, ruido granular, imágenes difusas, degradaciones de bloqueo, etc. Una vez provocadas estas distorsiones, sus magnitudes pueden medirse objetivamente y, haciendo uso de la experiencia adquirida con las evaluaciones subjetivas, esas mediciones podrían relacionarse entonces con alguna forma de cuantificación de la calidad del códec. En [Kobayashi, 1977] se dan ejemplos de estas mediciones para el caso de códecs intratrama y en [Hishiyama e Inoue, 1984] para el caso de códecs intercuadro. Es posible que resulte difícil correlacionar los factores de degradación con su interpretación por los observadores, en los sistemas de codificación intercuadro o en los sistemas de codificación que emplean algún tratamiento autoadaptable, porque pueden variar en cualquier momento, con el movimiento o adaptación del algoritmo de codificación. En [CCIR, 1982-86] se presenta un método de clasificación en tales casos. Según dicho método, la prueba de evaluación subjetiva utiliza primero escalas derivadas de pares de adjetivos opuestos (Método Diferencial Semántico) examinándose a continuación los resultados mediante análisis de componentes principales, para extraer los factores de degradación de la calidad de la imagen. Los resultados de la clasificación pueden verificarse aplicando un análisis de regresión múltiple que relaciona el factor con juicios subjetivos. En el Cuadro I se presenta una lista de factores de degradación de la calidad de la imagen.

Aunque estos métodos parecen convenientes para la evaluación de los códecs y ofrecen además un instrumento al diseñador, no resultan adecuados para referirse a la calidad de funcionamiento de un códec con imágenes reales por las siguientes razones:

- la composición compleja de las secuencias de imágenes reales no puede ser modelada satisfactoriamente mediante un número manejable de señales de prueba sintéticas;

- las degradaciones pueden ser de numerosos tipos y difíciles de clasificar por su naturaleza sutil (por ejemplo, una distorsión determinada puede ser visible en partes de una imagen con ciertas texturas que se mueven de una determinada manera);

- puede ser difícil definir mediciones objetivas significativas de las degradaciones (por ejemplo, en el caso de reproducción del movimiento). Debe observarse que la duración del periodo en el que se realizan las medidas objetivas debe corresponder a la ventana de observación proporcionada por la duración de la presentación en las pruebas subjetivas.

CUADRO I

Ejemplos de factores de degradación de la calidad de la imágen para un sistema digital, y medidas físicas correspondientes (unidades)

Factor de degradación de la calidad de la imagen	Medida física
- Imagen difusa	- Tiempo de establecimiento de la respuesta a la función escalón
- Actividad en los bordes	- Anchura de la fluctuación de fase de la respuesta a la función escalón
- Formación incorrecta de contornos	- Relación señal cresta a cresta/mínimo error de cuantificación cresta a cresta
- Ruido granular	- Equivalente analógico de la relación señal/ruido expresada en términos de señal cresta a cresta (S_{p-p}) a valor cuadrático medio del ruido (N_{rms})
- Efecto de "ventana sucia" (Dirty window effect)	- Amplitud máxima de ruido
- Sobrecarga en la pendiente temporal	- Tiempo de establecimiento de una transición em movimiemto
- Sacudidas	- Diferencia entre trama y cuadro en términos de bordes en movimiento
- Degradación tipo mosaico	- Por estudiar
- Bits erróneos	- Por estudiar

3.1.2 Métodos basados en material de imágenes naturales y error de codificación

Cabe asimilar las secuencias de imágenes naturales a una agrupación de cierto número de zonas diferentes, cada una con un contenido local distinto que hacen uso de códecs de longitud de palabra fija distintos de maneras distintas. Por ello el contenido de una secuencia de imágenes influye notablemente en la calidad percibida por el observador [Roufs y otros, 1989]. Es posible también cuando hay que comparar dos códecs en cuanto al contenido de secuencias de imágenes, determinar cuál parece mejor. Esto no sólo subraya la importancia de la elección de las imágenes de prueba para las evaluaciones subjetivas (véase el punto 9) sino también el hecho de que una medida objetiva de la calidad de funcionamiento de un determinado códec debe tener en cuenta el contenido de la imagen, si se pretende establecer una correlación entre los resultados de las evaluaciones subjetivas y objetivas.

Las formas más comunes de medición de la calidad objetiva se basan en el **error de codificación** de un códec; esto es, la diferencia entre una secuencia de imágenes de entrada y su salida decodificada. Esta misma señal de diferencia (a menudo amplificada) puede visualizarse como una secuencia de imágenes, lo que puede proporcionar una valiosa ayuda de diseño al especialista en códecs. Sin embargo, no debe utilizarse como material de evaluaciones subjetivas

3.1.3 Métodos basados en el error cuadrático medio normalizado

A menudo se utiliza como medida objetiva de la calidad de la imagen decodificada el error cuadrático medio de codificación. Se trata del promedio obtenido para cada muestra de imágenes de una secuencia, del cuadrado de los errores de codificación y está normalizado habitualmente con respecto a (al cuadrado de) la gama total de amplitudes de las muestras de imágenes. A veces se hace referencia al error cuadrático medio normalizado (NMSE) como un factor de ruido de codificación cuyo valor vendría dado por -10 log (NMSE). El uso generalizado de la medida del NMSE procede de su fácil expresión matemática pero ha de tenerse en cuenta con cierta cautela como una medida de la calidad de la imagen decodificada. No puede distinguir, por ejemplo, entre unos pocos grandes errores de codificación (que pueden ser tenidos como perturbando por un observador) y un gran número de pequeños errores de codificación (que pueden ser imperceptibles). Se ha tratado de ponderar la señal de error de codificación (tras una operación "log") antes de la evaluación del NMSE, con un filtro derivado de un modelo visual, y se ha conseguido mejorar la correlación con los resultados obtenidos en las evaluaciones subjetivas. El NMSE es un instrumento práctico y útil para el desarrollo de códecs, donde a menudo se necesita para comparar métodos de codificación que son muy similares (es decir, aquellos que utilizan variantes menores del mismo algoritmo y en los que los procesos de degradación pueden suponerse idénticos).

3.1.4 Métodos basados en modelos visuales

La sensibilidad del sistema visual humano al error de codificación en una determinada zona de una imagen se ve fuertemente influenciada por las características del propio material de imagen en esa zona. La incapacidad para tener esto en cuenta es la falla principal de la medida del NMSE. Un ejemplo útil, entre otros, que puede darse de esta influencia es el hecho conocido de que la sensibilidad de un observador al ruido de error de codificación se reduce

cuando el espectro de ese ruido coincide aproximadamente con el espectro de la imagen de "fondo". Estas propiedades del sistema visual son las que se aprovechan en el diseño de códecs, cuando se emplean experimentos subjetivos o datos psicovisuales para optimizar los parámetros del sistema.

Para mejorar la correlación entre las mediciones objetivas de la calidad de la imagen y el juicio de los observadores, es preciso desarrollar un modelo visual que pueda interpretar un error de codificación local en el contexto de la imagen de fondo y que pueda combinar todas esas evaluaciones locales hasta constituir una valoración global de la calidad. Este enfoque es aplicable tanto a códecs de longitud de palabra fija como a los de longitud de palabra variable, y se examina en el punto 3.2.3.

3.2 Códecs de longitud de palabra variable

Los códecs de televisión que necesitan reducir los datos de imagen de origen en un factor algo superior a dos, utilizan métodos basados en la codificación de longitud de palabra variable. La eficacia de estos códecs aumenta porque tienen la flexibilidad de atribuir dinámicamente bits de codificación a aquellas partes de una secuencia de imágenes en las que son más efectivos para preservar la calidad de la imagen decodificada. Esto lo pueden hacer los códecs de varias maneras sin que se implique necesariamente el uso de códigos de entropía de longitud variable.

3.2.1 Métodos basados en la utilización de señales de pruebas sintéticas

Debido a la flexibilidad de estos códecs, la degradación que introducen en cada muestra codificada depende no sólo de los valores de las muestras en la misma localización, sino también la historia de muestras anteriores uno o más cuadros hacia atrás. Esto significa que con códecs de longitud de palabra variable, ya sean del tipo intratrama o intercuadro, no tiene sentido tratar de caracterizar el códec intentando provocar distorsiones locales con señales de pruebas locales y hacer mediciones objetivas con ellas. Sin embargo, si los modos de adaptación de un códec de longitud de palabra variable pueden sostenerse artificialmente (lo que requiere acceso al interior de los mismos) cada modo puede caracterizarse por separado. El conocimiento de los conmutadores de adaptación del códec, cuando se le presentan escenas naturales, podría entonces utilizarse para determinar objetivamente su calidad de funcionamiento.

Es posible elaborar secuencias de prueba sintéticas en movimiento que lleven al códec al punto en el que produce distorsión visible, pero incluso si pudieran definirse las mediciones objetivas para caracterizar estas distorsiones (véanse las reservas al respecto en el punto 3.1.1), su interpretación sólo podría hacerse en el contexto de esa secuencia de prueba completa. Esto plantea interrogantes en cuanto a la medida en que esto es típico de escenas naturales y de si un diseñador de códecs tendría la posibilidad de optimizar el comportamiento de éstos para adecuarlos al material de prueba conocido.

3.2.2 Métodos basados en material de imágenes naturales y error de codificación

En cualquier evaluación de códecs de longitud de palabra variable es importante utilizar secuencias de imágenes naturales. Teniendo en cuenta la habilidad de estos códecs de dirigir la utilización de bits de codificación a lo largo de la imagen, habrá que considerar cuidadosamente el contenido de cada parte de la secuencia de imágenes cuando se enjuicie su criticidad (véase el punto 9). Se recomienda que cualquier evaluación objetiva se base en el error de codificación de un códec cuyas entradas sean un cierto número de imágenes de

prueba naturales. También puede aplicarse al error de codificación de los códecs de longitud de palabra variable el método del error cuadrático medio normalizado examinado en el punto 3.1.3, pero los resultados deben ser interpretados solamente por los especialistas, e incluso así, sólo como complemento de las evaluaciones subjetivas. De manera similar, sólo los especialistas en diseño de códecs deben efectuar las comparaciones objetivas entre códecs basadas en el NMSE y sólo cuando las técnicas a comparar difieren muy poco (es decir, sean variantes del mismo algoritmo) y cuando puede suponerse que los procesos de degradación son idénticos.

3.2.3 Métodos basados en modelos visuales

La principal desventaja de las medidas basadas en el NMSE es que no reconocen la fuerte influencia que el propio contenido de la imagen tiene en la sensibilidad de un observador a las degradaciones. Como se indicó en el punto 3.1.4, la optimización del diseño de los códecs implica el empleo de experimentos subjetivos y datos psicovisuales para ajustar la tolerancia del observador humano a la distorsión a las características de las zonas de imágen locales. Esto garantiza que cuando un códec de longitud de palabra variable reparta capacidad de bits de codificación (y por consiguiente reparta también las magnitudes de los errores de codificación) a lo largo de una imagen, lo haga de una manera que se adapte también a las características visuales. Por ello, cualquier método de evaluación objetiva debe incluir las propiedades del sistema visual humano, para producir resultados que estén en buena correlación con las valoraciones de calidad determinadas subjetivamente. La función de un modelo visual es la interpretación del error de codificación en el contexto de la imagen fuente en que éste se produce.

En el texto siguiente se supone que no se puede acceder a la parte interna de un códec. Si se puede obtener información sobre los modos de adaptación, los códecs de longitud de palabra variable también pueden evaluarse mediante el método de los factores de degradación (§ 3.1.1) según lo expuesto en [Inoue e Hishiyama, 1984].

En el desarrollo de un modelo visual, hay que incorporar dos niveles de conocimiento. El primero se refiere al grado de visibilidad de cualquier degradación arbitraria según su localización en la imagen, y el segundo determina cómo se deben combinar las visibilidades de todas las degradaciones para obtener una valoración global de la calidad. No obstante, sólo es necesario concentrarse en los modelos que tengan en cuenta las degradaciones características de los métodos de codificación digital; no hace falta considerar, por ejemplo, las distorsiones de naturaleza geométrica o semántica. Los modelos de la respuesta del sistema visual humano a las distorsiones causadas por las transmisiones de imágenes han centrado su atención en fenómenos situados en el umbral de visibilidad o cerca del mismo, lo que resulta adecuado para aplicaciones de televisión de alta calidad (véase, por ejemplo, [Sakrison, 1977]). Se conoce poco acerca de la modelación de la respuesta a distorsiones mayores.

En [Lukas y Budrikis, 1982] se presenta un importante estudio llevado a cabo por los autores en el que se detalla el diseño de un modelo visual de predicción de la calidad de las imágenes. En su trabajo examinan el desarrollo de este modelo y su comportamiento como predictor de la calidad subjetiva, desde un simple estimador basado en medidas de error no elaboradas, pasando por uno que modela el filtrado visual (no lineal), hasta uno que puede tener en cuenta las propiedades de enmascaramiento espacial y temporal de la visión. Como medios para este estudio se utilizaron los procesos de distorsión de cuantificación

uniforme, la codificación MICD, el ruido gaussiano aditivo y el filtrado paso bajo. El modelado del hecho observado de que los espectadores tienden a calificar las imágenes de acuerdo con el nivel de distorsión presente en el punto más deteriorado de las mismas y no según el valor medio de toda la imágen, resultó especialmente valioso para la obtención de una medida global de la calidad de una secuencia de imágenes. Más recientemente se ha desarrollado otro modelo visual aplicable a la codificación de imágenes digitales, [Girod, 1988] y [Zetzsche y Hauske, 1989].

La utilización de modelos visuales para la determinación objetiva de la calidad de la imagen en presencia no sólo de degradaciones de codificación digital sino también de degradaciones derivadas de otros procesos no lineales o adaptativos, es un tema muy prometedor. Desgraciadamente se le ha prestado poca atención, por lo que se encarece la presentación de Contribuciones sobre este asunto.

4. Evaluación objetiva de la calidad de imágen de los códecs en presencia de errores de transmisión

En un entorno de real transmisión, el enlace entre el codificador y el decodificador está sujeto a influencias que pueden corromper los datos transportados, por lo que una característica importante del decodificador es su respuesta a la presencia de esos errores de transmisión. En un códec diseñado cuidadosamente, esta respuesta tendrá la forma de distorsiones transitorias locales en la imagen decodificada, donde el número de esas distorsiones transitorias está relacionado con las estadísticas de error del canal, y su naturaleza, con el algoritmo de codificación de imagen empleado y con la criticidad de la secuencia de imágenes que se visualiza. Típicamente, el objetivo en las evaluaciones que implican errores de transmisión es obtener, para un códec dado, una representación gráfica de la degradación percibida por el observador en una gama de proporciones de errores.

Dentro de un decodificador existen varios niveles de tratamiento que determinan su respuesta a los errores de transmisión, algunos de los cuales pueden analizarse matemáticamente (o simularse por computador), mientras que otros requieren un cierto grado de evaluación subjetiva o bien un modelo objetivo de la respuesta del observador a las distorsiones transitorias.

La primera etapa de un análisis objetivo consiste en describir con la mayor precisión posible la forma en que ocurren los errores en un enlace práctico, esto es normalmente expresado mediante un modelo estadístico. En su forma más sencilla, dicho modelo supone que los errores se producen de manera aleatoria e independiente (distribución de Poisson), pero desde hace tiempo se sabe, a través de la observación empírica, que en realidad los errores aparecen agrupados o en ráfagas. Se han propuesto varios modelos para tener en cuenta este comportamiento; el más popular de ellos es el basado en la distribución tipo A de Neyman (véase, por ejemplo, [Jones y Pullum, 1981]). Mientras que la distribución de Poisson simple queda completamente definida con un sólo parámetro, la proporción media de bits erróneos, el modelo A de Neyman necesita que se cuantifiquen dos parámetros más relacionados con el grado de agrupación de los errores y la densidad de errores en cada agrupación. Todavía no se dispone de una recomendación para poder elegir de manera realista estos parámetros.

Los diseñadores de códecs, conocedores de que los errores de transmisión se presentan en ráfagas, incorporan a menudo un proceso de reordenación temporal de los bits transmitidos antes de que entren en el canal. Con ello se asegura la dispersión, por el mecanismo de reordenación inversa del

decodificador, de la ocurrencia en ráfagas de los errores de canal, haciéndolos así más receptivos al tratamiento por el subsiguiente sistema de corrección de errores. Este sistema de corrección de errores es capaz de corregir por completo cierto número de errores utilizando un complemento redundante de la capacidad de datos transmitidos pero queda cierta distribución de errores "residuales" que entran en el algoritmo de decodificación de imagen. Es posible calcular la distribución de errores residuales de un determinado códec y modelo de canal pero falta por evaluar el efecto que estos errores tienen en la imagen decodificada.

En [CCIR, 1986-90b] se sugiere que la calidad de un determinado códec en cuanto a errores de transmisión se juzgue en dos etapas: primero subjetivamente, para determinar la degradación debida a la característica transitoria de distorsión de ese códec, y después objetivamente, teniendo en cuenta la proporción de errores residuales obtenida por cómputo a partir de las consideraciones anteriores. En la actualidad no se dispone de evidencia experimental que sustente este enfoque, que podría ser no obstante, el primer paso hacia una medida totalmente objetiva si se pudiera caracterizar la respuesta del observador a diferentes transitorios del códec. Es importante tener en cuenta que algunos bits transmitidos son más sensibles a la corrupción que otros, lo que significa que la respuesta de un códec a un error residual de un solo bit puede variar mucho y puede depender además de la criticidad de la secuencia de imágenes de origen. Por ejemplo, en los códecs intercuadro la transitoria resultante de los errores residuales puede permanecer en partes estáticas de una secuencia de imágenes hasta que se tomen la medidas necesarias para eliminarla mediante refrescamiento. Finalmente, una característica de ciertos códecs que emplean codificación de longitud de palabra variable, es que pueden detectar algunas violaciones de la codificación causadas por errores de transmisión y utilizar ese conocimiento para tratar de ocultar las transitorias distorsionantes. Aunque no resulte satisfactorio para todos los errores, este proceso de ocultamiento mejora por lo general la calidad subjetiva de la imagen resultante, algo que debe tenerse en cuenta en cualquier evaluación objetiva de códec.

5. Evaluación subjetiva de la calidad de imagen de los códecs

Aunque se esté progresando al respecto, en la actualidad no se dispone de suficiente experiencia para dar detalles sobre métodos de evaluación objetiva de la calidad de imagen de los códecs. En materia de evaluación subjetiva, de la que existe mucha experiencia, se pueden hacer recomendaciones sobre condiciones de prueba y metodologías. Debe recordarse no obstante, al especificar objetivos de calidad o degradación, que los métodos existentes no pueden dar valoraciones subjetivas absolutas sino más bien resultados que están influidos en cierta medida por la elección de las condiciones de referencia y/o fijación. Pueden adoptarse las mismas metodologías para códecs de longitud de palabra fija y variable y para códecs de intratrama e intercuadro, aunque la elección de las secuencias de imágenes de prueba puede verse influenciada (véase el punto 9).

El método de evaluación más fiable para establecer un orden de jerarquía para los códecs de gran calidad consiste, en la actualidad, en evaluar todos los sistemas presentados al mismo tiempo y en condiciones idénticas. Las pruebas hechas independientemente, en las que se evalúan diferencias de calidad muy pequeñas, deben servir de guía más bien que de evidencia incuestionable de superioridad.

5.1 Evaluación de la calidad básica

Cuando se evalúa un códec para aplicaciones de distribución, esta calidad se refiere a las imágenes decodificadas después de un paso único a través de un par de códecs. En el caso de códecs de contribución, puede evaluarse la calidad básica después de varios códecs en serie, con el fin de simular así una aplicación típica de contribución.

5.1.1 Condiciones de observación y elección de los observadores

Se recomienda su conformidad con el punto 2.4 de la Recomendación 500 para televisión de resolución convencional y con la Recomendación 710 en el caso de códecs de TVAD.

5.1.2 Utilización de secuencias de imágenes de prueba

Se recomienda que en la evaluación se utilicen secuencias de al menos seis imágenes, más una adicional para los efectos de demostración antes del comienzo de la prueba. Las secuencias podrán tener una duración del orden de 10 segundos, pero debe señalarse que los evaluadores pueden preferir una duracion de 15-30 segundos [Inoue, 1988] [CCIR 1986-90c]. Deben variar entre moderadamente críticas y críticas en el contexto de la aplicación de reducción de velocidad binaria que esté en consideración (véase el punto 9).

5.1.3 Metodología de la prueba

Cuando la gama de calidades por evaluar es pequeña, lo que ocurrirá normalmente en el caso de códecs de televisión, la metodología de prueba a utilizar es la de doble estímulo con escala de calidad continua que se describe en la Recomendación 500 . La secuencia fuente original se utilizará como condición de referencia. En el Grupo Interino de Trabajo 11/4 se sigue debatiendo a propósito de la duración de la secuencia de presentación[CCIR, 1986-90d, e]. En pruebas recientes efectuadas por el Grupo Interino de Trabajo 11/7 en códecs para vídeo en componentes con relación 4:2:2 [CCIR, 1986-90a, f] con los resultados que figuran [CCIR, 1986-90g], se consideró ventajoso modificar la presentación con respecto a la que se da en la Recomendación 500 . Se utilizaron imágenes compuestas como referencia adicional para proporcionar un nivel de calidad inferior contra el cual juzgar el comportamiento del códec.

5.2 Evaluación de la calidad en el tratamiento posterior

Con esta evaluación se pretende facilitar la realización de apreciaciones sobre la idoneidad de un códec para aplicaciones de contribución con respecto a un determinado tratamiento posterior, por ejemplo la incrustación cromática, la cámara lenta o el zoom electrónico. La disposición de equipo mínima necesaria para tal evaluación consiste en un paso único a través del códec sometido a prueba, seguido del tratamiento posterior objeto de interés y a continuación, el observador. Sin embargo, puede ser más representativo de una aplicación de contribución el empleo de códecs adicionales después del tratamiento posterior.

5.2.1 Condiciones de observación y elección de los observadores

Véase el punto 5.1.1.

5.2.2 Utilización de secuencias de imágenes de prueba

Debido a las limitaciones de las posibilidades prácticas de tener que evaluar un códec con varios tratamientos posteriores, el número de secuencias de imágenes de prueba utilizadas puedeser como mínimo de tres, y una más disponible a efectos de demostración, porla imposición de tipo práctico de tener que evaluar probablemente un códec con varios tratamientos posteriores. La naturaleza de las secuencias dependerá de la tarea de tratamiento posterior que se estudie, pero debe variar entre moderadamente crítica y crítica en el contexto de reducción de la velocidad binaria de televisión y para el proceso que se considere. Las secuencias deberán tener una duración del orden de 10 segundos, pero debe señalarse que los evaluadores pueden preferir una duración de 15-30 s [Inoue, 1988] [CCIR, 1986-90c]. Para la evaluación de la cámara lenta puede servir una velocidad de visualización que sea la décima parte de la de origen.

5.2.3 Metodología de la prueba

La metodología de la prueba que debe utilizarse es la de doble estímulo con escala de calidad continua que se describe en la Recomendación 500. Sin embargo aquí la condición de referencia es la fuente sometida al mismo tratamiento posterior que las imágenes decodificadas. Si se considera ventajoso incluir una referencia de calidad inferior también ella deberá someterse al mismo tratamiento posterior. En las pruebas descritas en [CCIR, 1986-1990f], se ha introducido una ligera modificación de la presentación que se da en la Recomendación 500.

6. Evaluación subjetiva de la degradación de las imágenes de códecs debida a errores de transmisión

En el punto 4 se analiza en cierta medida cómo un decodificador digital maneja los errores de transmisión, con miras a considerar el modo de enfocar el análisis objetivo de la calidad de las imágenes. Una medida subjetiva útil puede ser la degradación determinada como función de la velocidad a la que se producen los errores de transmisión en el enlace entre el codificador y el decodificador. En la actualidad no se tiene conocimiento experimental suficiente de estadísticas ciertas de errores de transmisión, que permitan recomendar parámetros para un modelo que tenga en cuenta las agrupaciones o ráfagas de errores. En tanto no se disponga de esta información, pueden utilizarse los errores con la distribución de Poisson. En [CCIR, 1986-1990f] se dan algunos detalles sobre los corruptores de datos para aplicación en los niveles jerárquicos de transmisión de 34, 45 y 140 Mbit/s.

6.1 Utilización de secuencias de imágenes de prueba

Limitaciones de tipo práctico inducen a pensar que probablemente serán adecuadas tres secuencias de imágenes de prueba más una de demostración, puesto que hace falta explorar el comportamiento del códec con diversas proporciones de errores de transmisión. Cada secuencia debe tener una duración del orden de 10 segundos, pero debe señalarse que los evaluadores pueden preferir una duración de 15-30 s [Inoue, 1988] [CCIR, 1986-90c]. Estas deben variar entre moderadamente críticas y críticas en el contexto de reducción de la velocidad binaria de televisión (véase el punto 9).

6.2 Elección de las proporciones de errores

Deben elegirse al menor cinco proporciones de errores, pero preferiblemente más, con separación aproximadamente logarítmica y que abarquen la gama que provoca las degradaciones de códec desde "imperceptible" a "muy molesta".

6.3 Metodología de la prueba

Puesto que las pruebas abarcan la gama completa de degradaciones, el método de escala de degradación con doble estímulo (método UER) es el apropiado y el que debe utilizarse. El método se describe en la Recomendación 500.

6.4 Observación a propósito de la utilización de proporciones de errores muy bajas

Es posible que haga falta evaluar códecs con proporciones de errores de transmisión que provoquen transitorias visibles tan infrecuentes que no quepa esperar que se produzcan durante un periodo de secuencias de prueba de 10 segundos. El tiempo de presentación que aquí se sugiere es claramente inadecuado para tales pruebas.

Si es preciso grabar la salida de un códec en condiciones de proporción de errores bastante baja (lo que da lugar a un número pequeño de transitorios visibles en un periodo de 10 segundos) para montaje posterior en presentaciones de evaluación subjetiva, se debe tener la precaución de asegurarse de que la grabación utilizada es típica de la salida del códec observada en un intervalo de tiempo mayor.

7. Comparación subjetiva entre códecs

Cuando no hace falta una apreciación de la calidad o la degradación absolutas de un códec sino sólo su orden de jerarquía, o cuando se desea la confirmación del orden de jerarquía obtenido a partir de los resultados del método de doble estímulo, se debe utilizar el método de las comparaciones de pares de estímulos. Este método se da en el punto 4.2 de la Recomendación 500.

Tal como allí se describe, el método proporciona una comparación sensible y una manera de medir la relación entre pares de sistemas. En [CCIR, 1986-1990h] figura una extensión del método, para jerarquizar las calidades o las degradaciones de más de dos sistemas. En este enfoque, el orden de jerarquía global se deriva de la jerarquización de todos los pares posibles de secuencias de imágenes efectuada por los observadores.

El análisis se complica por el hecho de que un observador puede, por ejemplo, clasificar a la imagen A como mejor que la imagen B, y a la imagen B mejor que la C, pero también a la C mejor que la imagen A. Es lo que se denomina una "triada intransitiva". En [CCIR, 1986-1990h, i] se analiza el tratamiento estadístico de la transitividad para cada observador, así como el importante aspecto de la existencia estadísticamente significativa de un acuerdo sistemático importante entre observadores.

El número de presentaciones necesarias aumenta con el cuadrado del número de secuencias de imágenes de prueba y de códecs, lo cual representa una desventaja de este método que puede llegar a hacerlo impracticable.

8. Distorsiones en transmisiones mixtas analógicas y digitales

Hasta el momento, los problemas de especificación de la calidad de las imágenes se han considerado por separado para los sistemas analógicos y los sistemas digitales. Si se acepta la hipótesis de la independencia psicológica, de los fenómenos de degradación de la calidad de las imágenes mencionados en el punto 3.1, el enfoque descrito en ese punto puede ser aplicable también a los sistemas mixtos. Quiere decirse que esos fenómenos pueden clasificarse en uno de los tres grupos siguientes desde el punto de vista de la independencia psicológica:

a) degradaciones causadas únicamente por la sección analógica,

b) degradaciones causadas únicamente por la sección digital,

c) degradaciones causadas tanto por la sección analógica como por la sección digital (que pueden ser factores independientes en cada uno de los sistemas individuales).

Las degradaciones pertenecientes a los grupos a) o b) se considerarán como factores independientes, habiéndose ya propuesto al CCIR una ecuación para estimar la calidad global de las imágenes en este caso. Esta función de estimación puede aplicarse cuando existen varios factores psicológicos independientes entre sí que se dan simultáneamente.

Por otra parte, en el grupo c), en el que los fenómenos de degradación de la calidad de las imágenes en ambas secciones son tan similares que no pueden considerarse independientes, será necesario encontrar un nuevo método para asignar la degradación de la calidad de las imágenes tanto a la sección analógica como a la sección digital antes de aplicar la función de estimación mencionada anteriormente.

En la referencia [Inoue, 1987] se informa sobre un ejemplo de los resultados de la investigación en dicho caso. En ese trabajo se investigó una combinación de ruido aleatorio procedente de un sistema analógico y de ruido granular procedente de un sistema de codificación MICD intracuadro de longitud de palabra fija con el fin de demostrar que es posible sustituir una medida física en el sistema analógico con un valor corregido basado en las diferencias de sensibilidad visual.

9. Elección del material de imágenes de prueba para la evaluación de los códecs digitales

A lo largo de este Informe se ha hecho énfasis en la importancia de comprobar los códecs digitales con secuencias de imágenes que sean críticas en el contexto de la reducción de la velocidad binaria en televisión. Parece por ello razonable preguntarse en qué medida es crítica una secuencia de imágenes particular para un objetivo determinado de reducción de velocidad binaria, o si una secuencia es más crítica que otra. Una respuesta sencilla, aunque no especialmente útil, sería decir que "criticidad" significa cosas muy distintas para diferentes códecs. Por ejemplo, podría ocurrir que una imagen fija que contuviera muchos detalles resultase crítica para un códec intratrama mientras que para un códec intercuadro, que es capaz de aprovechar similitudes de cuadro a cuadro, esa misma escena no representaría ninguna dificultad. Algunas secuencias que emplean textura móvil y movimiento complejo resultan críticas

para toda clase de códecs, por lo que son las que más interesa generar o identificar. El movimiento complejo puede tomar la forma de movimientos que son predecibles para un observador pero no para los algoritmos de codificación, como por ejemplo un movimiento periódico tortuoso.

El análisis de posibles medidas estadísticas de criticidad de imagen [CCIR, 1986-1990j], por ejemplo mediante métodos correlativos, métodos espectrales, métodos de entropía condicional, etc., ha puesto de manifiesto una medida sencilla pero útil basada en una medición de entropía autoadaptable intratrama/intercuadro. Este método se empleó en la "calibración" de secuencias de imágenes propuestas para utilización en las pruebas de códecs para 34, 45 y 140 Mbit/s en el Grupo Interino de Trabajo 11/7, y demostró su utilidad para la selección de secuencias empleadas. La manera más sencilla de efectuar tales mediciones en secuencias de imágenes consiste en transferirlas a computadores de procesamiento de imágenes y someterlas a análisis por soporte lógico.

A continuación se dan algunas directrices de carácter general sobre cómo elegir material crítico, para el caso en que no se pueda acceder a las técnicas anteriores.

Códecs intratrama de longitud de palabra fija

Aunque es posible y válido evaluar estos códecs con imágenes fijas, se recomienda el empleo de secuencias móviles puesto que con ellas resulta más fácil observar los tratamientos del ruido de codificación y son más representativas de las aplicaciones de televisión. Si se emplean imágenes fijas en simulaciones de códecs por computador, se debe efectuar el tratamiento en toda la secuencia de evaluación, para preservar aspectos temporales de cualquier ruido de origen, por ejemplo. Las escenas elegidas deben contener el mayor número posible de los siguientes detalles: zonas estáticas con ciertas texturas y en movimiento (algunas con textura coloreada); objetos estáticos y en movimiento con bordes bruscos de alto contraste de diversas orientaciones (algunos con color); zonas estáticas uniformes semi-grises. Del conjunto de secuencias, al menos una debe presentar ruido de origen justamente perceptible y por lo menos una debe ser sintética (es decir, generada por computador) de modo que esté libre de imperfecciones de cámara tales como la abertura de exploración y persistencia de imagen.

Códecs intercuadro de longitud de palabra fija

Todas las escenas de prueba elegidas deben contener movimiento y el mayor número posible de los siguientes detalles: zonas con ciertas texturas y están en movimiento (algunas coloreadas); objetos con bordes bruscos de alto contraste moviéndose en dirección perpendicular a esos bordes y con diversas orientaciones (algunos coloreados). Del conjunto de secuencias, al menos una debe tener ruido de origen justamente perceptible y por lo menos una debe ser sintética.

Códecs intratrama de longitud de palabra variable

Se recomienda que estos códecs se prueben con material de secuencias de imágenes en movimiento, por las mismas razones que los códecs de longitud de palabra fija. Hay que tener en cuenta que debido a su codificación de longitud de palabra variable y su memoria intermedia asociada, estos códecs pueden distribuir dinámicamente la capacidad de bits de codificación a través de la

imagen. Así por ejemplo, si en la mitad de una imagen se presenta un cielo sin rasgos especiales que no necesita muchos bits para su codificación, se ahorra capacidad para otras partes de la imagen que pueden así reproducirse con calidad elevada, incluso si son críticas. La conclusión importante de todo esto es que si una secuencia de imágenes resulta crítica para un códec de este tipo, habrá que detallar el contenido de cada parte de la pantalla. Debe llenarse con textura en movimiento y estática, con tanta variación de color como se pueda y objetos con bordes bruscos de alto contraste. Al menos una secuencia del conjunto de prueba debe presentar ruido de origen justamente perceptible y por lo menos una debe ser sintética.

Códecs intercuadro de longitud de palabra variable

Este es el tipo de códec más complejo y el que necesita el material más exigente para forzarlo. No sólo hay que llenar cada parte de la escena con detalles como en el caso del códec intratrama de longitud de palabra variable, sino que esos detalles deben además estar en movimiento. Por otra parte, puesto que muchos códecs emplean métodos de compensación de movimiento, el movimiento a través de la secuencia debe ser complejo. Ejemplos de movimiento complejo son: escenas que emplean simultáneamente el zoom y las tomas con movimiento panorámico de la cámara; una escena que tenga como fondo una cortina agitada por el viento y en la que se aprecien sus detalles o su textura; una escena que contenga objetos que giran en un entorno tridimensional; escenas con objetos detallados que se aceleren a través de la pantalla. En todas las escenas debe abundar el movimiento de objetos con diferentes velocidades, texturas y bordes de alto contraste así como un contenido de color variado. Por lo menos una secuencia del conjunto de prueba debe tener ruido de origen justamente perceptible, al menos una debe tener movimiento complejo de cámara generado por computador a partir de una imagen fija natural (de modo que esté libre de ruido y persistencia de imagen de la cámara), y una secuencia cuando menos debe ser generada completamente por computador.

Las secuencias de prueba necesarias para las evaluaciones de tratamiento posterior están sujetas exactamente a los mismos criterios de criticidad. Sin embargo, es posible que resulte difícil cumplir con esos criterios en el caso de secuencias de primeros planos de incrustación cromática porque normalmente tienen una proporción importante de fondo azul sin rasgos característicos.

El Grupo Interino de Trabajo 11/7 ha preparado una amplia biblioteca de material de secuencias de prueba en formato de componentes con relación 4:2:2, que está grabada en cinta D1. En el Informe 1213 se dan detalles a propósito de estas secuencias, junto con los criterios con los cuales se prepararon (que pueden aplicarse a otras normas de televisión).

REFERENCIAS BIBLIOGRÁFICAS

Girod, B [1988] A Model of Human Visual Perception for the Reduction in Redundancy of Television Luminance Signals, PhD. Thesis University of Hannover 1988 (en alemán).

Hishiyama, K e Inoue, M [Septiembre, 1984] Physical Measures for Interframe Coded Picture Quality, Review of ECL, Vol 32, No 5.

INOUE, M. [Enero, 1988] - The influence of picture presentation period on subjective evaluation, Tech. Rep. of IEICE Japan, IE 87-105 (en japonés).

Inoue, M [Mayo, 1987] The Proposed Method for Noise Specifications to Mixed Analogue-Digital Video Transmission Systems, ITEJ, Vol 41, No 5 (en japonés).

INOUE, M. y HISHIYAMA, K. [1984] - Trade-off between information suppression effect and picture quality degradation with interframe coding, Review of the Electrical Communication Laboratories (Japón), Vol. 32, No. 5.

Jones, W J y Pullum, G G [Marzo, 1981] Error Performance Objectives for Integrated Services Digital Networks Exhibiting Error Clustering, Proc. IEE Conf. on Telecom Transmission, Londres, Publ no 193.

Kobayashi, Y [Mayo, 1977] Picture Quality Evaluation Method for Digitally Encoded Video Signals, Trans. Inst. Elect. Com. Engrs. Japan, Vol J60-B, No 5 (en japonés).

Lukas, F X J & Budrikis, Z L [Julio, 1982] Picture Quality Prediction Based on a Visual Model, IEEE Trans. Com., Vol COM-30, No 7.

ROUFS, J., DE RIDDER, H. y WESTERINK, J. [1989] - Perpetual image quality metrics, Institute for Perception Research, Eindhoven, Manuscript MS 692.

Sakrison, D J [Noviembre, 1977] On the Role of the Observer and a Distortion Measure in Image Transmission, IEEE Trans. Com., Vol COM-25, No 11.

Zetzsche, C y Hauske, G [Julio, 1989] Principal Features of Human Vision in the Context of Image Quality Models, Proc. IEE Int. Conf. on Image Processing and Applications, Warwick, Publ no 307.

Documentos del CCIR

[1982-86] : 11/26 (CMTT/31) (Japón).

[1986-90]: a. 11/498 (GIT 11/7); b. GIT 11/7-192 (Japón); c. 11/422 (Japón); d. GIT 11/4-154 (Japón); e. GIT 11/4-158 (Francia); f. GIT 11/4-182(Rev.2) (GIT 11/7); g. 11/498 (GIT 11/7); h. GIT 11/4-175 (Republica Federal de Alemania); i. GIT 11/4-156 (República Federal de Alemania); j. GIT 11/7-261 (Reino Unido).