UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

INGENIERÍA TÉCNICA DE TELECOMUNICACIÓN
ESPECIALIDAD: SONIDO E IMAGEN

PROYECTO FIN DE CARRERA

MEDIDAS DE CALIDAD SUBJETIVA EN SECUENCIAS DE VÍDEO

Autor: Carlos Esteban Baz Hormigos

Tutor: Manuel de Frutos López

Septiembre, 2009

Agradecimientos

No podría empezar este proyecto sin dar las gracias a mis padres y a mi hermano, que han hecho posible que esté a punto de ser ITT. Gracias por darme la oportunidad de estudiar lo que quería, por apoyarme, por llamar después de cada examen, por aguantar mis nervios, porque sé que también hacían los exámenes conmigo… Muchas gracias.

Gracias al resto de familiares que, aunque no enumere, son importantes para mí.

Gracias a mi tutor, Manuel de Frutos por la confianza depositada, por su ayuda y dedicación. Sobre todo, gracias por la disponibilidad a pesar de la distancia.

Quiero dar las gracias también a mis dos compañeros de piso, Aceituno y Carretero. Gracias por hacer estos años más fáciles, por todos los momentos buenos y malos que hemos pasado. Al final esas noches en vela haciendo prácticas han merecido la pena…

Por supuesto, gracias a los mejores amigos que uno puede tener: Abel, Alfredo, Carlos, David, Diego, Fernan, Nando, Rubén,… Gracias por haberme ayudado tanto estos últimos años de universidad, por esos viajes, esas reuniones los viernes en el burguer, esos consejos, por estar siempre ahí. Lo que me habría aburrido sin vosotros…

Gracias a todos mis compañeros de carrera, porque gran parte del camino que finaliza con este proyecto lo recorrimos juntos. Gracias por hacer más entretenidas las clases y laboratorios.

Agradecer también a todas las personas que, de forma desinteresada, participaron en la prueba subjetiva contribuyendo a la realización de este proyecto.

En definitiva, gracias a todos los que me han ayudado, acompañado, comprendido y aconsejado durante estos años. Espero seguir contando con vosotros.

Resumen

El objetivo principal de este proyecto es el desarrollo de una Medida Objetiva de Calidad Perceptual de secuencias de vídeo. Usando el algoritmo diseñado se implementará una aplicación que, de forma automática, proporcione una estimación de la calidad subjetiva de una secuencia a partir de su correspondiente referencia. Para ello se investigarán cuáles son los atributos visuales de mayor relevancia en la determinación de la calidad de vídeo, analizando los distintos componentes del Sistema Visual Humano.

Adicionalmente se estudiará el rendimiento de tres de las medidas objetivas más extendidas: la Medida de Calidad de Vídeo NTIA (modelo "general"), la Medida de Calidad de Vídeo Digital de Watson modificada y la Medida de Calidad de Vídeo VSSIM.

Se describirá, así mismo, la prueba experimental llevada a cabo como parte fundamental del proyecto. El propósito de la prueba, realizada según las correspondientes recomendaciones, es obtener valoraciones subjetivas a partir de un conjunto de observadores humanos. Esta información será utilizada no sólo para el diseño de la aplicación sino también para la posterior evaluación y comparación de los distintos algoritmos.

Abstract

The main goal of this project is to develop a Perceptual Quality Metric for video sequences. Using the designed algorithm, an automatic application will be implemented which will be able to predict subjective quality of a video sequence based on the corresponding reference. To do this, the most important visual attributes for determining video quality will be investigated, analyzing different properties on Human Visual System.

In addition, the performance of three widely-known objective video quality metrics will be studied. These metrics are the NTIA video quality metric ("general" model), a modified Watson's DVQ metric, and the VSSIM metric.

Moreover, it is described the experimental test performed to obtain subjective values from a group of human observers. The test, essential part of this project, was realized according to the corresponding recommendations. The achieved information will be used not only to design the application but also for the subsequent algorithm assessment and comparison.

Índice

Índice de figuras

Índice de tablas

1

Introducción

El vídeo digital está sujeto a varios tipos de distorsiones a través de los procesos de adquisición, compresión, procesado, almacenamiento, transmisión y reproducción. Por ejemplo, las técnicas de compresión de vídeo con pérdidas, que son las más usadas al reducir la capacidad necesaria en el almacenamiento o transmisión de los datos de vídeo, pueden degradar la calidad durante la cuantificación. Por otro lado, los flujos de bits de vídeo enviados a través de canales que introducen errores, como canales inalámbricos, pueden no ser recibidos correctamente. Además las redes basadas en conmutación de paquetes, como Internet, pueden causar pérdidas o retrasos considerables de los paquetes recibidos, dependiendo de las condiciones de la red y de la calidad de los servicios. Todos estos errores pueden traducirse en distorsiones en el vídeo recibido. Es, por lo tanto, imprescindible poder detectar y cuantificar las degradaciones en la calidad de vídeo que se producen en los sistemas, de forma que dicha calidad se pueda mantener, controlar, y posiblemente mejorar. Una medida efectiva de la calidad de vídeo es crucial para este propósito.

La manera más segura y fiable de determinar la calidad de una imagen o secuencia de vídeo es mediante la evaluación subjetiva, ya que los humanos serán los receptores finales en la mayoría de las aplicaciones. Es por ello que la puntuación media de opinión (MOS, *Mean Opinion Score*), que es una medida de calidad subjetiva obtenida a partir de un número de observadores humanos, ha sido reconocida durante muchos años como el método más fiable de evaluación de la calidad. Sin embargo, la obtención del MOS acarrea

una serie de desventajas obvias: el tiempo, los recursos necesarios, el coste y la imposibilidad de automatizar el proceso.

El propósito es alcanzar medidas de calidad objetivas con un alto grado de correlación con la calidad subjetiva, de modo que se pueda predecir la calidad perceptual de forma automática.

De forma general, una medida objetiva de la calidad de vídeo puede emplearse de tres maneras:

1. Se puede usar para monitorizar la calidad de imagen en sistemas de control de calidad. Por ejemplo, un sistema de adquisición de imágenes y vídeo puede usar la medida de calidad para ajustarse de forma automática con el fin de obtener la mejor calidad posible. Un servidor de vídeo puede examinar la calidad del vídeo transmitido en la red y controlar de esa forma el *streaming* de vídeo.

2. Se puede emplear para comparar sistemas de procesado de vídeo. Si hubiera disponible múltiples sistemas de procesado para una tarea específica, entonces una medida de calidad puede ayudar a determinar cual de ellos proporciona unos resultados con menor distorsión.

3. Puede ser integrada en el propio sistema de procesado de vídeo para optimizar el algoritmo y los parámetros de configuración.

Las medidas objetivas de calidad pueden ser clasificadas de acuerdo con la disponibilidad del vídeo original. Éste se considera sin distorsión o de calidad perfecta, y puede ser usado como referencia para comparar el vídeo distorsionado. La mayoría de los métodos propuestos en la literatura asumen que la secuencia original está completamente disponible. Aunque la expresión “calidad de vídeo” se usa a menudo por razones históricas, el término más preciso para este tipo de medidas sería *similitud* o *fidelidad* del vídeo. A menudo se denomina a estas técnicas *Full Reference* (FR) o de referencia completa. Sin embargo hay que tener en cuenta que en la mayoría de las aplicaciones prácticas, las secuencias de referencia no serán accesibles. Por lo tanto es muy deseable desarrollar técnicas de medida que puedan evaluar la calidad de forma ciega. Estas medidas, conocidas como *Non-Reference* (NR), o sin referencia, son muy complejas a pesar de que los

observadores humanos pueden evaluar sin dificultad y de forma efectiva y fiable la calidad de un vídeo sin ninguna otra referencia. Existe un tercer tipo de métodos en los cuales la señal de vídeo original no está completamente disponible. En su lugar, ciertas características se extraen de la señal de referencia y se transmiten al sistema de medida como información adicional. Estos métodos son conocidos como *Reduced-Reference* (RR), o de referencia reducida.

Este proyecto está centrado en los conceptos básicos, ideas y técnicas para la medida de calidad de tipo FR. Es importante destacar que un gran porcentaje de los modelos de medida FR propuestos inicialmente comparten una filosofía común basada en la detección de error. Posteriormente, se introduce una nueva forma de pensar en el problema de la evaluación de calidad de imagen y vídeo y se proporcionan algunos resultados preliminares de un nuevo método FR basado en la distorsión estructural.

1.1 Estado del arte

A lo largo de muchos años de investigación en el procesado visual, uno de los problemas fundamentales ha sido la ausencia de una métrica universalmente aceptada que proporcione una medida de calidad con unas prestaciones aceptables en un amplio rango de situaciones. En general, para obtener la estimación más precisa de la calidad de imagen o vídeo se han utilizado pruebas subjetivas. Sin embargo, como se ha comentado anteriormente, debido a los inconvenientes que dichas pruebas acarrean se hace necesario el uso de métricas de calidad objetivas. De éstas, las más conocidas, y que se han empleado principalmente (y se siguen empleando) son simples medidas matemáticas.

Una imagen o un vídeo cuya calidad desea ser evaluada puede modelarse como suma de una señal perfecta más una señal de error. Se puede asumir que la pérdida de calidad está directamente relacionada con la potencia de la señal de error. Por lo tanto, una forma natural de evaluar la calidad de la imagen es cuantificar el error entre la señal distorsionada y la original. Las medidas más comunes de este tipo forman un conjunto de parámetros de calidad objetiva conocidos como *Medidas de Distorsión Comparativa* y explican el hecho

de que grandes valores de error corresponden a una mala calidad de imagen. Algunas de estas medidas son el error cuadrático medio, la desviación típica o la relación señal a ruido (se analizarán con más detalle en el capítulo 3).

Sin embargo, este tipo de métricas han sido muy criticadas por no correlacionar bien con la medida de calidad percibida [1-4]. Su mayor inconveniente es que sólo son eficientes cuando los errores se comportan como ruido adicional que no está correlacionado con la señal. Es decir, no se puede distinguir entre pocos errores de gran amplitud (los cuales son molestos para los observadores) y muchos errores con baja amplitud (que desde un punto de vista subjetivo pueden ser imperceptibles). Estas medidas de distorsión comparativa pueden ser más eficientes desde el punto de vista del dominio frecuencial cuando se aplican a imágenes previamente filtradas con filtros de modelado del Sistema Visual Humano. Con ello se puede dar una mejor estimación de la importancia del error para el ojo humano.

De hecho, en las últimas tres o cuatro décadas, se ha realizado un gran esfuerzo para desarrollar métodos de medida objetivos (la mayoría de tipo FR), que incorporen medidas de calidad perceptual considerando las características del SVH. La primera medida cuantitativa de calidad fue propuesta en 1982 por Lukas y Budrikis [5]. Este campo de investigación ha aumentado su actividad en gran medida desde entonces, incluso alguno de los modelos desarrollados está hoy en día disponible comercialmente. En octubre de 1997 se formó el Grupo de Expertos en Calidad de Vídeo (VQEG, por sus siglas en inglés) con el fin de desarrollar, validar y estandarizar los nuevos métodos de medida objetivos en calidad de vídeo. Aunque la fase I [6,7] para evaluación de calidad en secuencias de televisión alcanzó un resultado limitado, el VQEG continúa su trabajo con la fase II de evaluación de calidad en televisión y multimedia, incluyendo también técnicas RR y NR. Por lo tanto, la investigación de estas técnicas de evaluación de calidad de vídeo está todavía en proceso. De momento, únicamente se ha conseguido éxito limitado en evaluaciones de sofisticados sistemas de medida FR basados en el SVH bajo estrictas condiciones de prueba y un amplio rango de tipos de distorsiones [6,7].

La implementación de estas métricas es en muchos casos compleja, los algoritmos son bastante elaborados y con un coste computacional alto. El grado de complejidad depende de las características visuales que se contemplen dentro de la métrica y el modelo escogido para las mismas. La investigación de evaluación de calidad comenzó centrada en imágenes

estáticas; posteriormente se aplicaron dichos conocimientos a las secuencias de vídeo. Inicialmente consistía en aplicar las técnicas existentes a cada uno de los fotogramas del vídeo para después realizar algún tipo de combinación que proporcionara un valor único. Poco a poco, a los algoritmos se fueron incorporando técnicas que tenían en cuenta los aspectos temporales en la evaluación de calidad.

Sin embargo, los primeros modelos presentaban limitaciones al no tener en cuenta aspectos de alto nivel del SVH como extracción de características, procesos cognitivos, reconocimiento de patrones, atención visual, etc. Por ello, las medidas posteriores han intentado incorporar características del procesado del SVH de alto nivel. En otros casos se han buscado nuevas métricas que, de alguna forma, simulen ciertas características de la percepción sin llegar a implementar un modelo visual completo.

En definitiva este es un área con distintos frentes de investigación abiertos puesto que aún no hay una métrica estándar aceptada para la evaluación de la calidad percibida. Además, el estudio del comportamiento de las métricas ya existentes es complicado puesto que la validación de las mismas ha de hacerse mediante pruebas subjetivas sobre un conjunto de secuencias dadas con lo que, en cualquier caso, su evaluación es parcial.

1.2 Objetivos

El objetivo final de este proyecto es la implementación y evaluación de una medida objetiva de calidad perceptual de vídeo. Adicionalmente se realizará una comparación de los resultados obtenidos con los algoritmos actualmente más extendidos.

Como se ha comentado, la obtención del MOS acarrea una serie de desventajas obvias (tiempo, recursos necesarios, coste, imposibilidad de automatizar el proceso). Por ello se persigue alcanzar una medida de calidad objetiva con un alto grado de correlación con la calidad subjetiva, de forma que se pueda predecir la calidad percibida de forma automática.

Por lo tanto, se hace necesaria la realización de una prueba experimental cuyo propósito es obtener las valoraciones subjetivas que servirán tanto para el diseño de la medida de calidad, como para examinar el rendimiento de los algoritmos analizados.

En la figura siguiente se muestra el proceso de desarrollo necesario para alcanzar los objetivos. Para el diseño de la medida de calidad propuesta, se usa un conjunto de secuencias (formado por pares de secuencia original y su correspondiente versión distorsionada) que se emplean en la prueba subjetiva. Estos mismos pares de secuencias se procesan con la medida objetiva y se comparan los resultados con los del experimento mediante análisis estadísticos. A partir de los análisis se fijan los parámetros de la medida objetiva para obtener la mayor correlación posible con los valores de MOS.

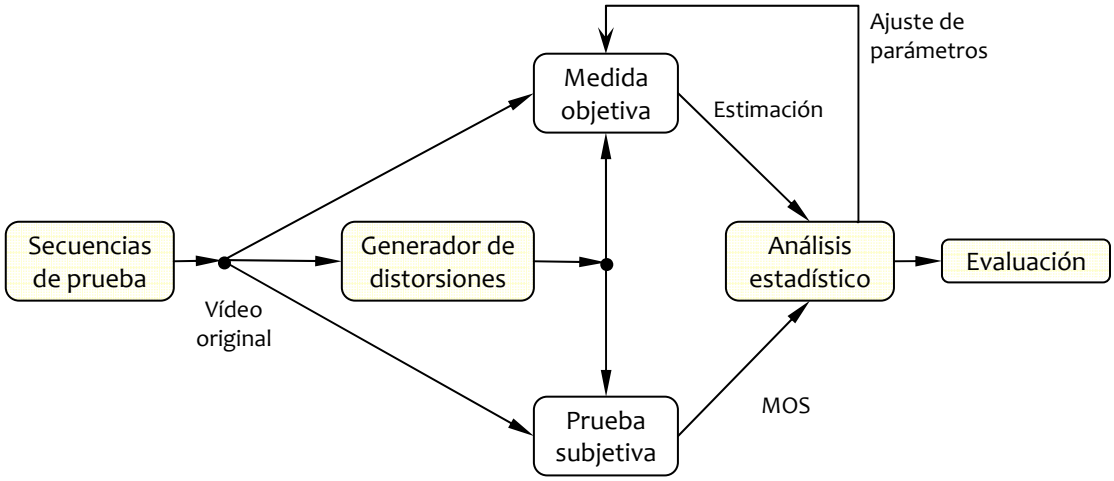


Fig. 1.1 Proceso de desarrollo de la medida de calidad de vídeo propuesta

Posteriormente, haciendo uso de los resultados de la prueba subjetiva se realizarán análisis estadísticos con el resto de algoritmos analizados para comprobar la eficacia de cada una de las medidas.

1.3 Estructura de la memoria

El proyecto cuenta con siete capítulos, mediante los cuales se describe el proceso seguido para llegar a la implementación y evaluación de la medida objetiva de calidad perceptual de vídeo.

En este primer capítulo se tratan los motivos fundamentales que han suscitado la creación del proyecto, los objetivos propuestos para su desarrollo y la estructura que se utilizará en la memoria.

En el segundo capítulo se realiza una descripción del Sistema Visual Humano y sus propiedades que será de gran importancia a la hora de comprender y diseñar medidas acordes con la calidad percibida por el observador final.

En el tercer capítulo se hace una revisión de las posibles distorsiones en secuencias de vídeo y se introducen las dos filosofías principales: detección de error y distorsión estructural.

En el cuarto capítulo se analizan con detalle los algoritmos seleccionados. Se estudia su estructura básica así como las técnicas empleadas en cada uno de ellos.

En el quinto capítulo se realiza una descripción completa de la medida de calidad de vídeo propuesta. Se analizan los objetivos, la estructura general, las técnicas en las que se basa y los módulos adicionales. Por último se introduce la aplicación software realizada.

El sexto capítulo detalla la prueba subjetiva llevada a cabo, así como los resultados obtenidos. A continuación se realizan análisis estadísticos comparando la eficacia de los métodos analizados con respecto a la medida propuesta.

Por último en el séptimo capítulo se hace una reflexión sobre el trabajo desarrollado, problemas encontrados, etc. Se exponen algunas ideas sobre trabajos futuros posibles a partir de las conclusiones obtenidas.

Posteriormente se incluyen tres anexos. En el primero se muestra una estimación de los costes del proyecto. El segundo incluye el manual de usuario de la aplicación desarrollada. El tercero recoge los resultados detallados de la prueba subjetiva y de los algoritmos analizados.

Finalmente se indican las referencias bibliográficas que se han consultado para realizar el proyecto.

2

Sistema Visual Humano

Para desarrollar y entender medidas objetivas acordes con la calidad percibida por el observador final se han de tener en cuenta los aspectos más importantes tanto fisiológicos como psicológicos del SVH.

En el estudio de la visión se pueden diferenciar claramente dos etapas. Por un lado, se realiza un procesado a bajo nivel que no utiliza información de experiencias pasadas ni razonamiento. Por otra parte, existe un procesado de alto nivel que utiliza recursos como la atención y la memoria, con una complejidad mucho mayor y, por tanto, bastante más difícil de modelar.

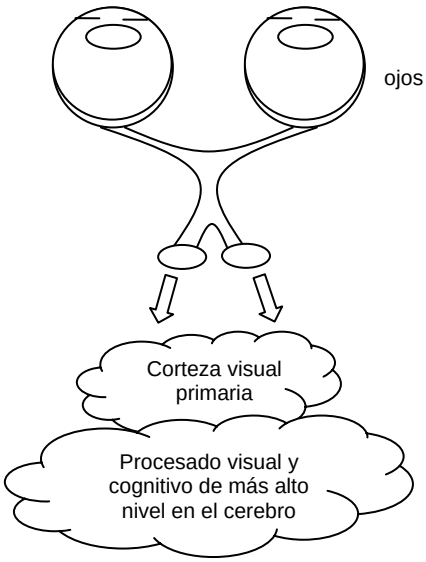


Fig. 2.1 Diagrama esquemático del sistema visual humano

La mayor parte de las métricas de calidad basadas en las características del SVH se centran en el procesado a bajo nivel [8-10]. Todavía no se conoce con exactitud cómo el cerebro humano extrae información de más alto nivel cognitivo en etapas posteriores de la visión.

En este capítulo se pretende analizar las características de comportamiento del SVH, comprendiendo su funcionamiento y determinando, en la medida de lo posible, su respuesta. Esta será la base sobre la que posteriormente se podrán modelar medidas de calidad perceptual.

2.1 Anatomía y fisiología visual

2.1.1 El ojo

El ojo es el órgano del SVH en el que comienza el procesamiento visual y cuyo funcionamiento básico consiste en recoger y enfocar los estímulos visuales en forma de luz que provienen de los objetos del entorno en su superficie posterior, transformando energía luminosa en eléctrica. Es sensible a las radiaciones electromagnéticas con longitudes de onda comprendidas entre los 400 y los 780 nm aproximadamente.

La figura inferior muestra una sección del globo ocular, con una forma prácticamente esférica y de unos 20 mm de diámetro.

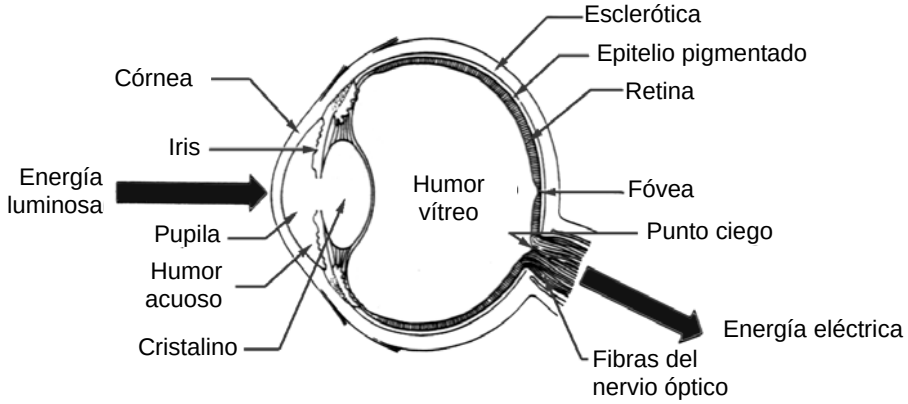


Fig. 2.2 Sección vertical del ojo

En la figura, se pueden apreciar los componentes principales que constituyen el ojo, que van a ser enumerados a continuación, explicando su funcionamiento y características más relevantes.

La esclerótica es el tejido duro, blanco y fibroso que conforma la parte exterior del globo ocular. Es opaca y se conoce comúnmente como el blanco del ojo. La parte delantera y central del ojo es la córnea. Se trata de una membrana transparente y dura cuya función principal consiste en refractar la luz. Además, la córnea tiene una cierta curvatura que contribuye a dirigir la luz y concentrarla en una pequeña apertura redonda (de 2 a 8 mm de diámetro) situada en el centro del ojo y que permite que la luz pase al interior del mismo, la pupila. Ésta a su vez está rodeada por el iris, un músculo circular que permite regular la cantidad de luz que entra en el ojo, de forma que si la intensidad luminosa es elevada se contraerá haciendo menor el diámetro de la pupila y viceversa. La parte anterior del iris es la que contiene el pigmento visible que caracteriza el color de ojos.

Tras el iris se encuentra el cristalino. Está formado por una serie de capas concéntricas con distintos índices de refracción que actúan como una lente flexible enfocando la luz en el fondo del ojo. El cristalino absorbe cerca del 8% de la luz visible del espectro. Esta absorción se incrementa en zonas del infrarrojo y del ultravioleta, pudiendo llegar a dañarse el ojo por un exceso de radiación en estas frecuencias. Entre la córnea y el cristalino se encuentra una sustancia líquida, clara y transparente denominada humor acuoso.

La retina recubre la parte interior del ojo y es en ella donde se enfoca la luz incidente, que se convierte en señales nerviosas mediante células sensibles a la luz. Entre el cristalino y la retina se encuentra el humor vítreo que es una sustancia gelatinosa transparente e incolora. El humor vítreo llena todo el espacio entre el cristalino y la retina y ocupa alrededor de 2/3 del volumen ocular. Para poder enfocar objetos cercanos y lejanos es necesario que el ojo humano cambie la forma del cristalino. Este proceso, que se denomina acomodación, es controlado mediante un grupo de músculos situados alrededor del iris y sucede prácticamente en tiempo real. Una vez realizada esta acomodación, la luz se proyecta en la retina y es transformada en impulsos eléctricos mediante dos tipos de fotorreceptores que reciben el nombre de conos y bastones.

Los conos se distribuyen fundamentalmente por la zona central de la retina denominada fóvea, son sensibles al color y responsables de la visión fotópica (en condiciones normales de iluminación). Proporcionan una visión en detalle. Existen diferentes tipos de conos, que corresponden con las tres diferentes longitudes de onda a las que son más sensibles. Los conos L, conos M y conos S (en función de la longitud de onda a la que se sitúan sus picos de sensibilidad: *Long*, *Medium* y *Short* o larga, media y corta) dividen la imagen proyectada en la retina en tres flujos de información. Estos flujos pueden ser vistos como las componentes roja, verde y azul del estímulo visual, aunque la aproximación es bastante burda. Cada ojo posee entre 6 y 7 millones de conos y su umbral de visibilidad se sitúa alrededor de 1μL (micro-lumen).

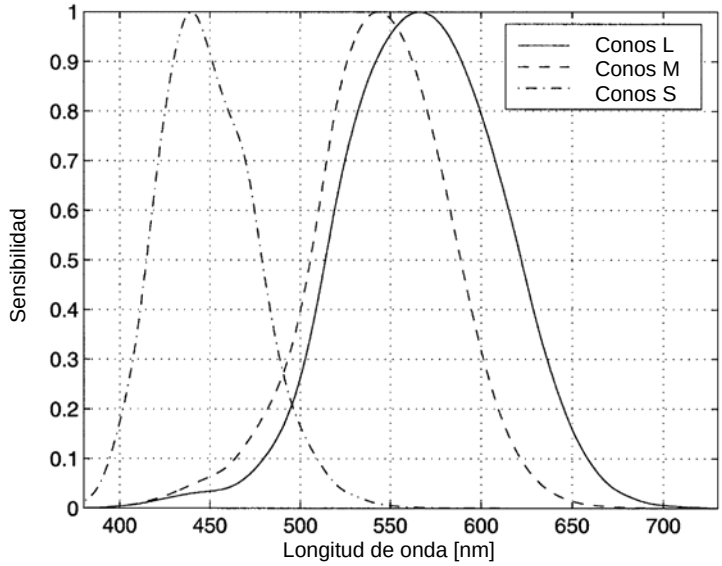


Fig. 2.3 Sensibilidad espectral normalizada de los tres tipos de conos

Los bastones proporcionan información general del campo visual. Son sensibles a valores muy bajos de iluminación con umbrales de 1nL, de tal forma que hacen posible la visión nocturna o con poca iluminación pero sin información de color. Son responsables, por tanto, de la visión escotópica y de ahí que normalmente sean ignorados en el modelado. El número de bastones es de 75 a 150 millones.

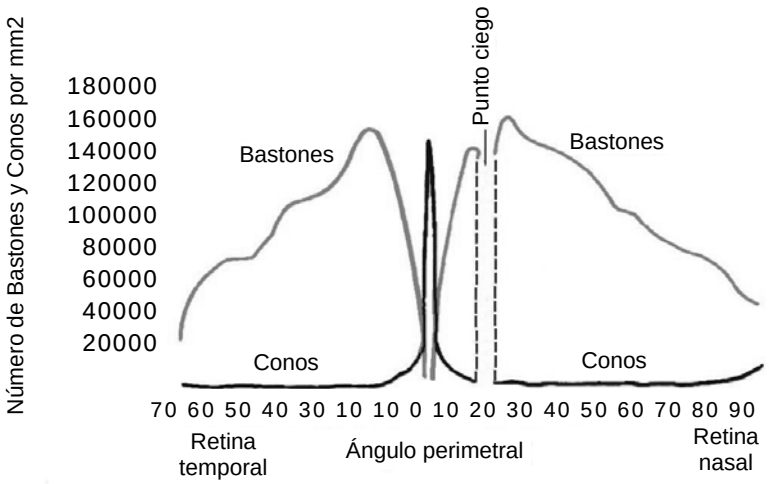


Fig. 2.4 Distribución de conos y bastones en la retina

Como puede apreciarse en la figura 2.4 la distribución de estos fotorreceptores en la retina es poco uniforme y más o menos simétrica respecto a la fóvea. Como consecuencia, tenemos un amplio ángulo de visión y una resolución espacial elevada a nivel local que se resuelve con el continuo movimiento del ojo enfocando los objetos de interés en la fóvea.

El disco óptico se caracteriza por ser la única zona de la retina en la que no hay fotorreceptores, constituyendo, por tanto, un punto ciego. De aquí parten los vasos sanguíneos que riegan la retina y además en esta zona convergen los axones formando el nervio óptico que parte con la información que será enviada al cerebro.

Como se ha comentado, todos estos elementos ópticos enfocan el estímulo visual en la retina, pero en este proceso se puede producir un desenfoque de la imagen debido a limitaciones e imperfecciones. El desenfoque es paso bajo, modelado normalmente como un sistema lineal e invariante en el espacio caracterizado por una función de distorsión, también conocida como función de dispersión de punto (PSF, *Point Spread Function*).

2.1.2 Campos receptores de la retina

La figura inferior muestra una sección de la retina con las células que la forman y su conexionado. Como puede observarse, entre las células ganglionares (células más avanzadas del procesamiento visual) y los fotorreceptores existen otros tres tipos de células: horizontales, amacrinas y bipolares.

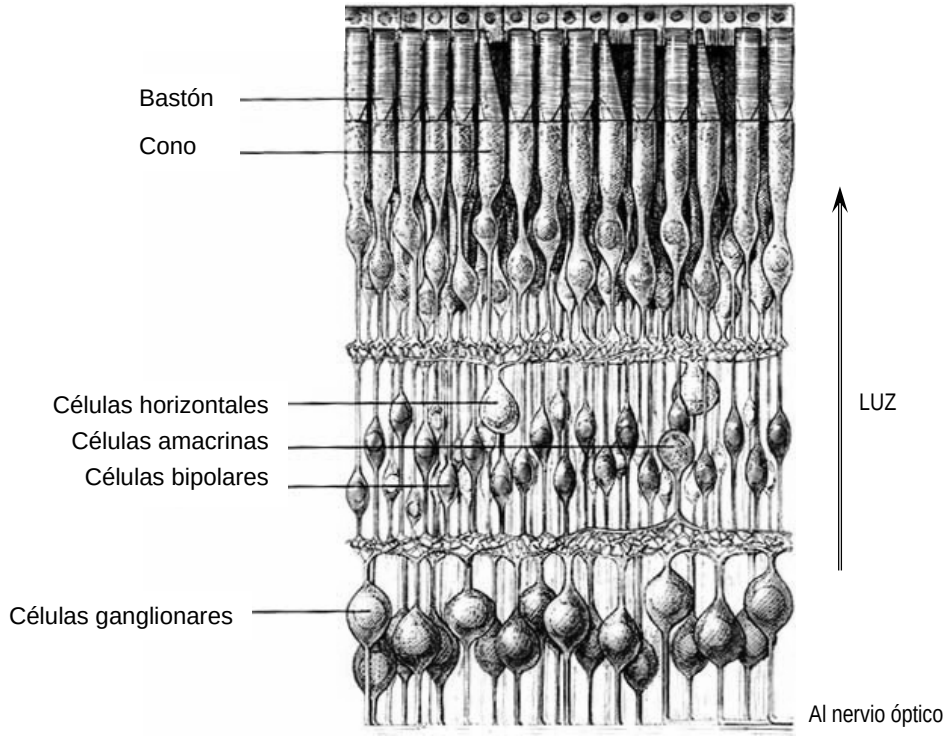


Fig. 2.5 Sección de la retina (longitud aproximada de ¼ de mm)

- Las *células horizontales* conectan bastones y conos vecinos. Tienen un efecto inhibidor sobre las células bipolares.
- Las *células bipolares* conectan células horizontales, bastones y conos con las células ganglionares. Las células bipolares pueden tener salidas excitadoras o inhibidoras.
- Las *células amacrinas* transmiten señales de las células bipolares a las células ganglionares o lateralmente entre diferentes neuronas. Se conocen más de 30 tipos diferentes en cuanto a su anatomía y pueden desarrollar otras funciones entre las que se encuentran respuestas específicas a objetos en movimiento.

- Las *células ganglionares* recogen información de las células bipolares y amacrinas. Hay alrededor de 1,6 millones de este tipo de células en la retina. Sus axones forman el nervio óptico que sale del ojo y lleva la señal de salida de la retina hasta otros puntos de procesado en el cerebro.

Las interconexiones entre estas células dan lugar a un importante concepto en la percepción visual, los campos receptores. El campo receptor visual de una neurona se define como el área de la retina en el que al incidir la luz sobre ella, la respuesta de dicha neurona se modifica.

Las células ganglionares de la retina tienen un campo receptor de tipo centro-periferia, casi simétrico circularmente. Se han identificado dos clases: células ganglionares de centro encendido (excitadoras u "on") y células de centro apagado (inhibidoras u "off"). Las células ganglionares de centro encendido se activan cuando el centro de su campo receptor se encuentra excitado y la periferia del mismo no. En las células de centro apagado sucederá a la inversa. De esta manera, si las dos partes del campo receptor reciben el tipo de estimulación que necesitan, sus efectos se suman y se alcanza un nivel de activación máximo en la ganglionar, pero si el estímulo produce efectos opuestos en el centro y en la periferia, las dos regiones compiten entre sí y la célula ganglionar correspondiente se mantiene casi inactiva. Esta interacción es conocida como inhibición lateral. La inhibición lateral es uno de los procesos más importantes en la explicación de muchos fenómenos perceptivos, y en concreto, este particular funcionamiento hace que ya en las primeras etapas de la visión se pierda prácticamente la información de intensidad luminosa en niveles absolutos, en favor de una percepción del contraste.

De la descripción que se ha hecho hasta ahora se podría pensar que los campos receptores forman un mosaico de pequeños círculos retinianos, pero no es así, los campos receptores de células ganglionares vecinas se solapan como puede apreciarse en la figura 2.6 b. Un fotorreceptor puede influir sobre varias células ganglionares estando situado en el centro del campo receptor de algunas y en la periferia de otras. Por tanto, excitará a varias células a través de sus centros si son células de centro encendido y a través de sus periferias si son células de centro apagado.

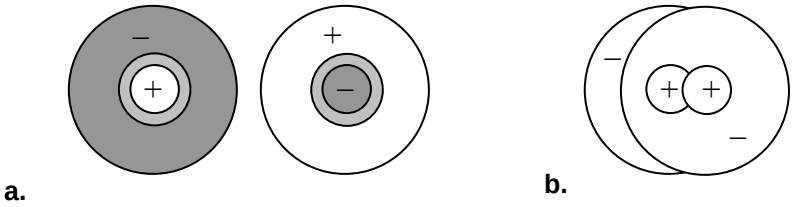


Fig. 2.6 Sección Campos receptores. **a.** Campos receptores de célula ganglionar de centro encendido y centro apagado respectivamente. Entre el centro y la periferia, existe una pequeña zona de respuesta mixta **b.** Campos receptores de dos células ganglionares vecinas.

Además, los campos receptores difieren en tamaño de una célula ganglionar a otra. En concreto, los centros de los campos receptores son más pequeños en la fóvea y se hacen mayores a medida que nos alejamos de ella. Esto, nuevamente, explica el hecho de que nuestra agudeza visual, la capacidad de distinguir objetos pequeños, sea mayor en la fóvea. El tamaño del campo receptor está pues relacionado con la frecuencia espacial de forma que células con campos receptores grandes responderán a bajas frecuencias espaciales y viceversa. De acuerdo con esto se puede hacer una nueva clasificación de las células ganglionares en: magno (10% células) que son las que poseen campos receptores grandes y parvo (90% células) con un campo receptor pequeño. Las células magno responden al movimiento y a objetos grandes, son rápidas y de alta sensibilidad al contraste. Las parvo son lentas y están implicadas en la detección del color y del detalle.

En conclusión, los mensajes que el ojo envía hacia el cerebro a través del nervio óptico tienen poco que ver con intensidad absoluta ya que las células ganglionares no responden bien a cambios de luz difusa. Lo que la célula señala es el resultado de una comparación entre la cantidad de luz que impacta en un cierto punto de la retina con la cantidad media de luz que ilumina su periferia inmediata. Esto permite ver el mismo objeto en condiciones de iluminación completamente distintas. Por ejemplo, si leemos un periódico a plena luz del sol o en una habitación poco iluminada lo que vemos son letras negras sobre un fondo blanco. Sin embargo, si tuviésemos en cuenta la luz que reflejan, las letras negras a plena luz reflejarían más que el papel blanco en la habitación.

Se hace evidente, por otro lado, que la retina es mucho más que un dispositivo que convierte luz a señales neuronales. La información visual sufre en ella un pre-procesado exhaustivo antes de ser enviada a otras partes del cerebro.

2.1.3 Vías visuales. Integración de la información en la corteza

La vía visual del cerebro humano es el camino que sigue la información visual desde el ojo (células ganglionares) hasta que llega a la corteza visual primaria.

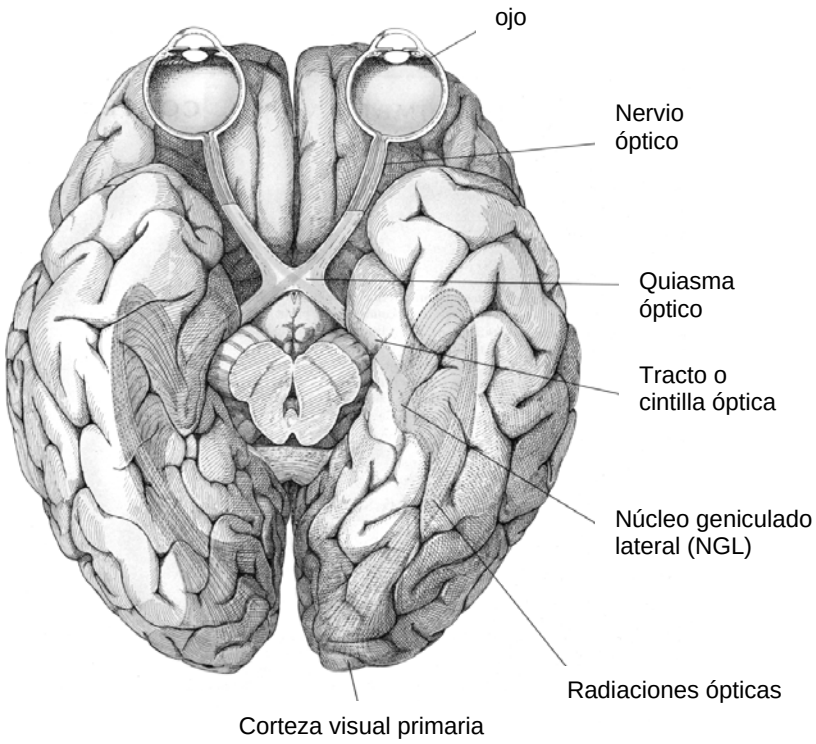


Fig. 2.7 Vías ópticas hacia la corteza visual

Las fibras que salen del ojo formando el nervio óptico llegan sin interrupción al quiasma óptico y a partir de ahí se dirigen a distintas zonas del cerebro. La mayoría (80% aproximadamente) envía la información a través del tracto óptico al núcleo geniculado lateral, el resto es información para el control de funciones de movimiento ocular y reflejo pupilar a la luz, así como para procesos de sincronización de ritmos biológicos.

En el quiasma óptico se cruzan los nervios ópticos de forma que a cada núcleo geniculado lateral le llega información procedente de ambos ojos, pero correspondiente a la mitad opuesta del campo visual. Es decir, al núcleo geniculado lateral derecho le llega toda la información visual correspondiente al campo visual izquierdo. El campo visual es la zona del espacio que puede ser percibida para una posición fija del ojo. El campo visual

izquierdo se refiere a la zona del espacio a la izquierda del punto al que estemos mirando. De esta forma, cada hemisferio cerebral se encarga de la mitad opuesta del entorno.

Cada cuerpo geniculado lateral se encuentra dividido en 6 capas. Las dos capas más interiores (capas magnocelulares) reciben información casi exclusivamente de células ganglionares de tipo magno. Las otras cuatro capas (capas parvocelulares) reciben información principalmente de células parvo. No obstante, una célula de una capa cualquiera recibe información de un solo ojo.

La información visual sale de los cuerpos geniculados laterales mediante unas bandas anchas denominadas radiaciones ópticas que la encaminan hasta la corteza visual primaria. La corteza visual primaria es una capa de células de unos 2mm de grosor en la parte occipital del cerebro. Contiene aproximadamente 200 millones de células frente a los 1,5 del cuerpo geniculado lateral. Estas células fueron clasificadas por Hubel [11] en células simples y complejas.

Las células simples tienen campos receptores de tipo "on" y "off" pero con patrones de excitación e inhibición como los mostrados en la figura 2.8 a. Estímulos que cubran una mayor área excitadora darán lugar a mayores respuestas y los que cubran regiones excitadoras e inhibidoras a la vez producirán respuestas menores puesto que hay una cancelación mutua. Estas células poseen, además, tres o cuatro geometrías diferentes con todas las orientaciones y posiciones dentro del campo visual posibles. Son, por tanto, las primeras células que presentan sensibilidad a la orientación.

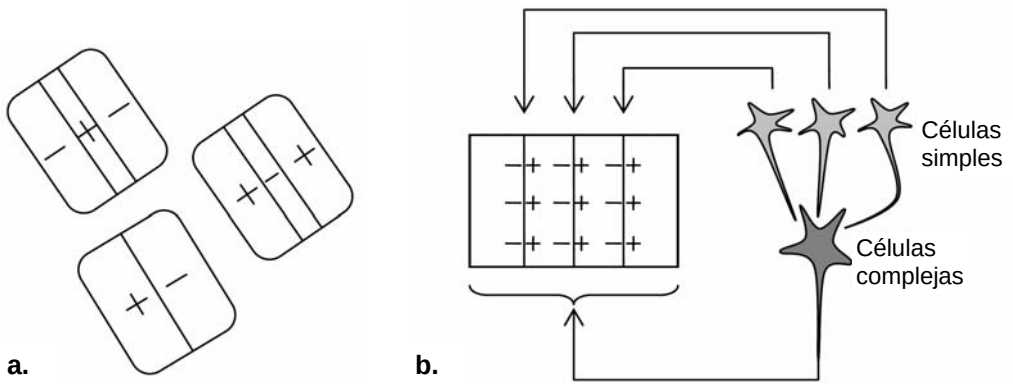


Fig. 2.8 Campos receptores. **a.** Mapas típicos de campos receptores de células simples. **b.** Posible campo receptor y diagrama de conexión de célula compleja

Finalmente, y por primera vez en todo el proceso de percepción, aparecen células en la corteza visual primaria que reciben información de ambos ojos; es decir, hay una convergencia binocular que permitirá construir una única imagen de la escena. En general todas las gradaciones de dominancia ocular relativa están presentes (desde células monopolizadas por el ojo izquierdo pasando por células con igual influencia de ambos ojos hasta células con respuesta al ojo derecho).

La corteza visual primaria se estratifica en 6 capas. Las más densas son la 4C y la 6 mientras que la 1 apenas contiene células nerviosas y está formada en su mayor parte por axones. En la figura 2.9 se pueden observar las principales conexiones efectuadas entre el cuerpo geniculado lateral y las distintas capas de la corteza visual primaria, así como las conexiones entre capas y con otras regiones del cerebro.

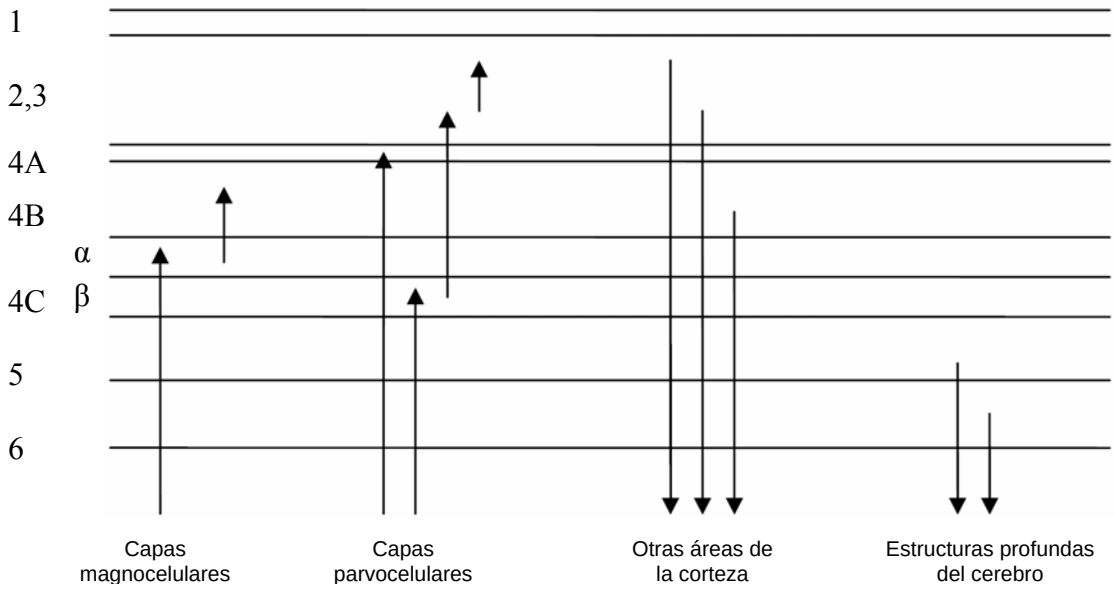


Fig. 2.9 Estructura e interconexiones en la corteza visual primaria

Las células que forman la corteza visual primaria se encuentran dispuestas en columnas verticales relacionadas con la orientación del estímulo. Además, se produce también una agrupación de células sensibles a una determinada frecuencia espacial. De esta forma se configuran detectores de frecuencia espacial. Se debe tener en cuenta que hay también otras capas de la corteza (desde la 2 a la 5) que participan en el proceso visual realizando diversas tareas. Así, por ejemplo, la zona 4 estaría relacionada con el procesamiento del color.

En conclusión, el sistema visual humano que se ha descrito es modular y paralelo. Se podrían diferenciar básicamente tres etapas. La primera de ellas sería la óptica (equivalente a un enfoque), la segunda la retiniana (transducción de la señal luminosa en determinados impulsos eléctricos) y finalmente el procesado cerebral.

2.2 Propiedades de la visión

A continuación se analizan algunos de los mecanismos más relevantes de la visión. Se pretende realizar una caracterización del SVH en cuanto a su sensibilidad y respuesta al estímulo.

2.2.1. Adaptación visual. Sensibilidad a la intensidad luminosa

El rango de niveles de intensidad luminosa que se presentan en la naturaleza y a los que el SVH se tiene que adaptar es muy amplio, alrededor de 1010. Sin embargo, simultáneamente, el SVH es capaz de discriminar unos pocos niveles. La adaptación se realiza para un nivel de intensidad luminosa I0 dentro de todos los posibles del rango. Para este I0, se pueden discriminar unos 50 valores distintos de intensidad. Por tanto, habrá un nivel por debajo del cual no se distingan otros niveles, que se verían como negro y de igual forma sucederá con la percepción del blanco. A medida que varía la intensidad que recibe el ojo, éste va definiendo los umbrales de negro y blanco y recogiendo los distintos valores intermedios entre dichos umbrales.

La adaptación proporciona la capacidad de ver en condiciones de muy distinta iluminación. A pesar de ello, hay que tener en cuenta que la percepción visual va a variar mucho dependiendo de los niveles de luz existentes. Así, como ya se ha comentado, en condiciones de baja iluminación se detectan muy bien las diferencias de luminosidad, pero la distinción del color y del detalle es pobre y sucederá justo al contrario en condiciones de elevada luminosidad. Además, el proceso de adaptación no es instantáneo, siendo más costosa temporalmente la adaptación de luz a oscuridad.

En el sistema visual humano podemos distinguir tres mecanismos de adaptación:

- La variación mecánica de la apertura de la pupila, controlada por el iris. El diámetro de la pupila puede variar entre 1,5 y 8 mm, lo que corresponde a una variación de 30 en la cantidad de luz que entra al ojo. Este mecanismo de adaptación responde en cuestión de segundos.
- Procesos químicos en los fotorreceptores. Este mecanismo está presente tanto en conos como en bastones. En condiciones de mucha luz, la concentración de partículas fotoquímicas de los receptores disminuye, reduciendo su sensibilidad. Este mecanismo es muy potente (cubre de 5 a 6 órdenes de magnitud), sin embargo es muy lento; por ejemplo, la adaptación a oscuridad total puede llevar en torno a una hora.
- Adaptación a nivel neuronal. Este mecanismo afecta a las neuronas de todas las capas de la retina, que se adaptan a la luz incidente incrementando o decrementando su señal de salida. Esta adaptación es menos potente que la adaptación química, pero es mucho más rápida.

2.2.2. Sensibilidad al contraste

Por la descripción fisiológica previa del SVH se sabe que los mecanismos básicos de la visión trabajan con el contraste. Experimentos psicofísicos que evidencian los fenómenos visuales relativos a la percepción de contraste pueden verse en la figura 2.10. En la figura de las bandas de Mach se muestra el fenómeno de inhibición lateral que se produce en los campos receptores y a pesar de que cada barra representa un nivel constante se aprecian diferencias subjetivas en los bordes de separación de las bandas. Se puede observar también un ejemplo de contraste simultáneo, en el que se muestra un rectángulo interior del mismo nivel de intensidad pero que se percibe más claro u oscuro según el nivel de luminosidad del rectángulo exterior.

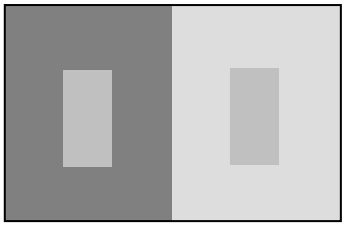
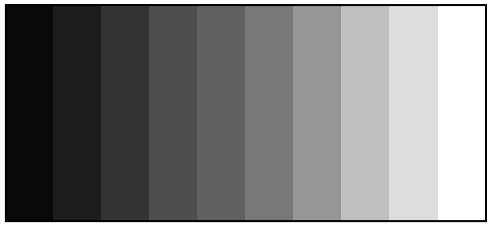
a.  **b.** 

Fig. 2.10 Fenómenos relativos al procesamiento de contraste. **a.** Contraste simultáneo **b.** Bandas de Mach

Por tanto, como el SVH es sensible al contraste y no a la luminancia en términos absolutos, es de gran importancia la definición que se haga del contraste. Lo que podría parecer en un primer momento sencillo es, en realidad, bastante complicado y su dificultad aumenta cuando se quiere definir el contraste para imágenes complejas. El objetivo sería encontrar una medida de contraste que se corresponda con la percepción visual que se tiene del mismo.

El problema se ha resuelto en primer lugar con imágenes simples (patrones) como por ejemplo funciones sinusoidales o discos de luminancia. Para este tipo de imágenes se utilizan dos definiciones de contraste diferentes. Por una parte, Michelson definió el contraste de este tipo de imágenes como:

$$C_{Michelson} = \frac{L_{máx} - L_{mín}}{L_{máx} + L_{mín}}$$

donde $L_{máx}$ y $L_{mín}$ son los valores máximos y mínimos de luminancia. Weber propone otra definición para medir el contraste local de un objeto simple sobre un fondo uniforme:

$$C_{Weber} = \frac{\Delta L}{L}$$

donde ΔL es el incremento o decremento de la luminancia del objeto comparada con la luminancia del fondo L.

La diferencia entre ambas definiciones es clara y se hace más obvia si expresamos el contraste de Michelson en la siguiente forma:

$$C_{Michelson} = \frac{\Delta L}{L + \Delta L}$$

donde $\Delta L = (L_{máx} - L_{mín})/2$ y $L = L_{mín}$. Estas dos medidas de contraste no coinciden a pesar de que pueden proporcionar valores parecidos para estímulos simples de bajo contraste y ambas definen el contraste como una relación adimensional de la variación de luminancia respecto a la luminancia media de fondo. Sin embargo para contrastes más elevados se obtienen dos medidas de contraste diferentes y lo que es más, estas dos definiciones ni siquiera comparten el mismo rango. Mientras que el $C_{Michelson}$ tiene un margen de valores de *0* a *1*, el C_{Weber} varía entre *–1* e ∞.

Además, se plantean otros problemas derivados del hecho de que tanto el $C_{Michelson}$ como el C_{Weber} se definen para imágenes simples y de forma global. En imágenes más complejas sería más apropiada una definición local del contraste puesto que la luminancia (y con ella también el contraste) varía en toda la imagen. Por tanto, las fórmulas anteriores no serán válidas para el cálculo de contraste en imágenes complejas como las que se tratan en cualquier modelo de visión. Estas medidas tampoco tienen en cuenta que la sensibilidad al contraste del SVH es dependiente de la frecuencia, sobre todo si nos encontramos en condiciones umbrales de visibilidad.

Para solucionar estos problemas, Peli [12] propone un cálculo del contraste por bandas de frecuencia. Es un contraste local y limitado en banda que tiene en cuenta el nivel medio de luminancia local para la obtención de un valor en cada punto. Puesto que el cálculo del contraste se efectúa por bandas es necesario disponer primero de una versión filtrada de la imagen.

Otras posibles métricas de contraste para imágenes complejas realizan una modificación del contraste de Michelson adaptándolo al filtrado previo. Daly [13] propone un cálculo global y otro local del contraste. Tanto en este caso, como en todos los anteriores, se debe tener en cuenta que la definición que se hace del contraste debe ser consecuente con la implementación del modelo visual que se esté llevando a cabo así como con el objetivo que va a tener este modelado.

Por otro lado, es importante destacar que la sensibilidad al contraste es mayor en la fóvea y decrece al alejarse de ella para cada frecuencia espacial. Esta disminución es más rápida para frecuencias altas, de tal forma que la función de sensibilidad al contraste se transforma en paso bajo en la periferia de la retina.

2.2.3. Sensibilidad en frecuencia

El estudio de la anatomía y fisiología del SVH ha permitido que se conozca que la percepción del estímulo depende de la frecuencia del mismo. Tradicionalmente una de las caracterizaciones más importantes que se hacen del SVH consiste en la máxima frecuencia que es capaz de detectar o diferenciar y que se denomina agudeza visual. Este dato

proporciona únicamente un límite y es insuficiente si se quiere conocer la variación en la detectabilidad de un estímulo en función de su frecuencia espacial. Con este objetivo se empezó a medir la función de sensibilidad al contraste (CSF, *Contrast Sensitivity Function*) que refleja la sensibilidad o capacidad de detección del SVH a estímulos de distinta frecuencia. La CSF fue determinada por primera vez en 1956 por Schade pero su uso no se generalizó hasta que las técnicas de Fourier empezaron a utilizarse en visión en los años 70.

Puesto que lo que se pretende medir es la respuesta del SVH a distintas frecuencias, el procedimiento de medida consistiría en mantener un estímulo de contraste constante y variarlo en frecuencia para ver cómo el SVH atenúa cada frecuencia. Sin embargo, esto resulta imposible puesto que no podemos medir el contraste de la imagen percibida por lo que es necesario realizar el procedimiento justo al contrario. La medida se realiza para cada frecuencia variando el contraste del estímulo de entrada y lo que mantenemos constante es la salida, es decir, la imagen percibida.

Un procedimiento experimental para calcular la CSF consistiría, por ejemplo, en ir reduciendo el contraste de un estímulo sinusoidal en el que la luminancia media se mantiene constante hasta que se alcanza el umbral (el estímulo deja de ser visible). El inverso de este valor umbral es el valor de la sensibilidad para esa frecuencia. Al realizar este proceso para las diferentes frecuencias obtenemos la CSF. Se pueden hacer varias consideraciones sobre este método de medida. En primer lugar, la CSF así calculada sería válida para estímulos simples, no para imágenes complejas. Aunque cualquier estímulo complejo pueda ser analizado como una serie de estímulos sinusoidales, el problema reside en que el SVH es no lineal y por ello, la respuesta obtenida varía y no se puede calcular como una combinación lineal. En segundo lugar, este método es de cálculo de umbrales de detección y es válido sólo en condiciones umbral, la forma de la CSF para condiciones por encima del umbral no se corresponde con esta medida. Para condiciones por encima del umbral habría que llevar a cabo otro tipo de pruebas, con el inconveniente de que son experimentos menos estables y de mayor dificultad de medida, por lo que no suelen utilizarse.

Existen muchas medidas experimentales de la CSF, la mayoría de ellas para estímulos monocromáticos. En [16] se dispone de un resumen de algunas de las CSF existentes y los parámetros que se han tenido en cuenta en la medida. La forma típica de la función puede

observarse en la figura 2.11 y es la de un filtro paso banda con un pico que se sitúa entre los 4 y 8 ciclos/grado para niveles de iluminación fotópicos. Se produce una fuerte atenuación en altas frecuencias debida a factores ópticos como imperfecciones en la córnea y cristalino, difracciones en la pupila… Además, el muestreo espacial que realizan los fotorreceptores de la retina impone un límite a partir del cual no se detectan frecuencias espaciales mayores (agudeza visual). La atenuación que aparece en bajas frecuencias se debe a las interacciones inhibitorias de centro-periferia que se producen en los campos receptores. La medida del valor de la CSF en frecuencia cero es imposible para el método de medida que se ha explicado con anterioridad y según los parámetros que se hayan escogido al realizar los experimentos para la obtención de la CSF se dará una frecuencia mínima a partir de la cual se empieza a medir.

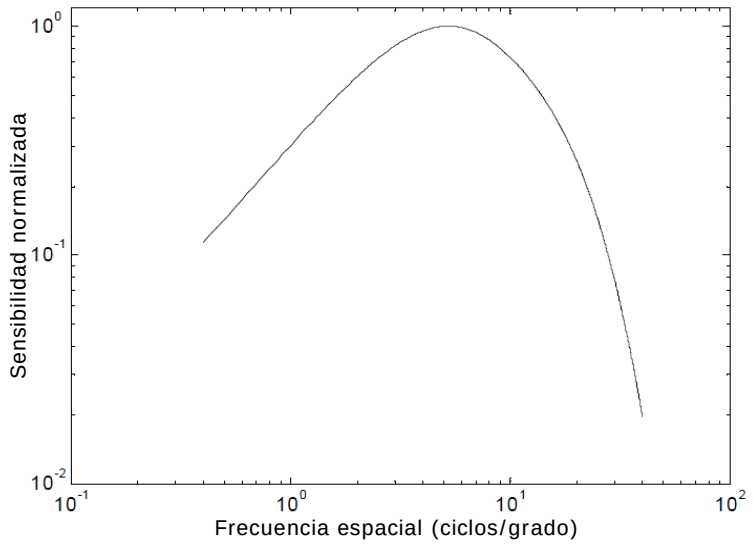


Fig. 2.11 Función de sensibilidad al contraste normalizada

Es necesario destacar que aunque se ha expuesto únicamente y de forma muy sencilla un método de medida de la CSF umbral, éste, por supuesto, no es el único. Existe una gran variedad de métodos entre los que se encuentran los tests impresos [16], que aunque no son muy precisos tienen como ventaja su gran sencillez. Además, dentro de la medida de CSF umbral se pueden seguir distintos métodos psicofísicos. En definitiva, la elección de uno u otro método dependerá de la aplicación en la que se vaya a utilizar la CSF y de la precisión o las características especiales que se requieran para la misma. Así, un

posible uso de la CSF sería, por ejemplo, la detección y seguimiento de algunas afecciones visuales que tienen un efecto determinado en la forma de la curva, y que presenta como ventaja que es un método no invasivo y de detección temprana. En este caso bastaría con medir la CSF para el paciente bajo observación. Otra posible utilización de la CSF, sería dentro de un modelado del SVH genérico, lo que reviste una mayor complejidad puesto que no existe una CSF "universal".

2.2.4. Enmascaramiento

El enmascaramiento es el fenómeno por el cual disminuye la visibilidad de una señal en presencia de otra que esconde o enmascara a la primera. No se debe olvidar que el método utilizado para la obtención de los umbrales de visibilidad trabaja con estímulos muy simples de una sola frecuencia y sobre campos de intensidad constante. Desde este punto de partida, el enmascaramiento es un paso más de complejidad en el modelado del SVH que ha sido observado y estudiado ampliamente por fisiólogos y psicofísicos.

El objetivo principal será conocer cómo varían los umbrales de visibilidad del estímulo cuando se encuentran en el entorno cambios espaciales y también temporales de luminancia. Se dan entonces dos fenómenos de enmascaramiento claramente diferenciados:

- *Enmascaramiento temporal:* asociado a la reducción visual que se produce sobre el estímulo cuando está rodeado de cambios temporales de luminancia. El estudio de este tipo de enmascaramiento es complicado debido a que habría que tener en cuenta las características de movimiento del ojo, es decir, hay una dependencia con la velocidad de movimiento retiniana. Además, habría que considerar si el estímulo en movimiento es seguido o no por el ojo, lo que depende del interés particular del observador, siendo esto muy difícil de analizar.

  Por otra parte, está demostrado que cuando hay un cambio de escena, la resolución espacial puede ser reducida drásticamente sin que se llegue a percibir siempre que se restablezca la resolución original en un periodo breve de tiempo (en torno a 100 ms).

- *Enmascaramiento espacial:* está asociado a los cambios espaciales de luminancia que se producen alrededor del estímulo. Recibe también el nombre de enmascaramiento de contraste ya que el efecto del enmascaramiento será menor en regiones uniformes que en zonas de gran contraste. Así, si se considera cualquier imagen natural a la que se le añade un ruido aleatorio uniforme, la visibilidad del ruido en los bordes y texturas que presenta la imagen será menor que en zonas uniformes.

  El fenómeno del enmascaramiento espacial se estudia mediante tests psicofísicos en los que se evalúa la influencia del contraste, de la frecuencia y de la orientación tanto del estímulo (enmascarado) como de la señal de fondo (máscara). Las señales utilizadas suelen ser sinusoides y de Gabor. Estos experimentos han demostrado que el umbral de detección que proporcionaba la CSF va incrementando a medida que el efecto del enmascaramiento se hace mayor, lo que ocurre cuando el contraste de la señal que enmascara se aumenta.

El comportamiento típico se muestra en la figura 2.12. C_T representa el valor de contraste para que el estímulo sea visible y C_M el contraste de la señal que enmascara. Para un valor pequeño de C_M, el umbral de visibilidad del estímulo C_T viene dado por la CSF y se representa como C_0. Este valor se mantiene hasta que se alcanza un punto en el que $C_M = C_0$. A partir de ahí un incremento en C_M hace que se eleve el umbral de detección, es decir, se necesitaría aumentar el contraste del estímulo para que fuera visible. La línea discontinua de la figura ha sido observada en algunos de los experimentos psicofísicos y proporciona una medida por debajo del valor dado por la CSF, sin embargo, este efecto sólo se observa cuando la señal y el ruido están en fase.

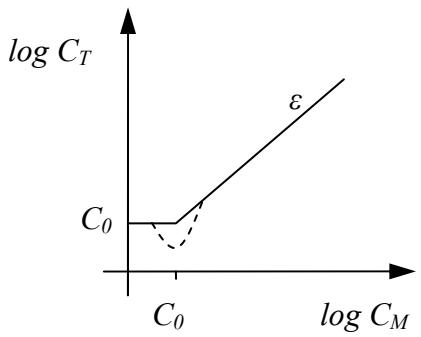


Fig. 2.12 Curva de umbrales de elevación causados por enmascaramiento espacial.

El valor de la pendiente ε depende tanto de la complejidad y conocimiento del estímulo que se tengan, como del tipo de señal utilizada para el enmascaramiento. Si se conocen datos de la orientación, tamaño, frecuencia y localización del estímulo será más fácil reconocerlo y por tanto la pendiente de la curva será menor. El valor de ε varía entre 0,6 y 1 dependiendo de que el tipo de señal utilizada para el enmascaramiento sea una sinusoide o un ruido limitado en banda. Además, aparece un efecto de aprendizaje. Si a un observador se le muestra el mismo patrón de ruido repetidas veces, el valor de ε decae puesto que será capaz de reconocerlo con un valor menor de contraste.

La frecuencia espacial y la orientación también influyen en el grado de enmascaramiento, que será máximo cuando se produzca una coincidencia entre la frecuencia de la señal que enmascara y el estímulo enmascarado.

2.3 Movimientos oculares y atención visual

Los ojos se fijan a la cabeza mediante tres pares de músculos que le permiten orientarse en dirección vertical, horizontal o circular. La cantidad de información con la que debe tratar el SVH y que se encuentra en el entorno visual hace que sean necesarios mecanismos selectores de información. Así, la atención puede ser considerada como un conjunto de redes de áreas neuronales que llevan a cabo operaciones específicas de procesamiento de información. Dentro de estas operaciones específicas se encontraría la detección o selección de objetivos dentro de la escena, que se corresponderá con determinados movimientos oculares.

Si se analiza la forma que tienen los ojos de explorar el entorno visual se van a encontrar tres tipos de movimientos oculares, atendiendo a la funcionalidad de los mismos:

1) Movimientos para el mantenimiento de la mirada: son aquellos que compensan el movimiento de la cabeza o de los objetos para que permanezca la mirada fija sobre el objeto.

2) Movimientos para el desplazamiento de la mirada: permiten pasar la atención de un objeto a otro. Fundamentalmente, se dan tres tipos: sacádicos, persecuciones o de seguimiento y vergencias.

3) Movimientos de fijación o micromovimientos: trémores, microsacádicos y fluctuaciones.

Si se considera el proceso seguido al visualizar una imagen estática se observa un primer paso en el que se fija la posición de un objeto o detalle por un breve periodo de tiempo (100-500 ms) para, posteriormente, realizar un salto a una nueva posición, este salto se denomina movimiento sacádico. Por tanto, la exploración visual que se hace de una escena estática no es un movimiento continuo como cabría esperar, sino una sucesión de saltos de un punto de interés a otro. Además, cuando se tiene fijada una posición, los ojos no permanecen quietos, se realizan movimientos muy pequeños y constantes. Estos micromovimientos son continuados y de dirección aleatoria y su importancia radica en que permiten que se vean los objetos estacionarios. De hecho si se mantuviese una imagen estable en la retina, en vez de obtener una mayor agudeza visual (se evitaría el emborronamiento asociado al movimiento de la imagen que producen los movimientos oculares), lo que sucede es que los receptores de la retina se saturan y la imagen desaparece.

La particular forma de ver imágenes estáticas dando saltos de un punto de la imagen a otro hace que se ponga de manifiesto la importancia que tiene el estudio de la atención visual. De hecho, aunque es posible prestar atención a una parte de la escena sin que se produzca un movimiento ocular, el caso contrario es imposible: cualquier movimiento sacádico viene precedido por un cambio de atención a esa posición concreta. Así, parece poco probable que en la visualización de una imagen, se haga un salto al vacío, es decir, que en un movimiento se fije un punto que corresponde con fondo o con un punto que carece de cambios abruptos de luminancia.

El estudio de la atención visual controlando los puntos de parada de la visión resulta de gran interés y proporciona gran cantidad de información acerca de los objetos y detalles que son relevantes para el observador. Los factores que influyen en la atención visual [18] son muchos y variados, entre ellos se pueden destacar el tamaño, color, contraste, forma, bordes, texturas y por supuesto el movimiento.

Los movimientos de persecución o de seguimiento se producen de forma coordinada con ambos ojos, su finalidad es la de seguir estímulos visuales que se desplazan lentamente. Su velocidad oscila entre 1-30 º / s. A priori no son movimientos voluntarios, tratan de estabilizar la imagen visual en movimiento sobre la retina; sin embargo es posible ejercer control sobre los mismos mediante entrenamiento.

Alternativamente a los movimientos de persecución, las vergencias implican el movimiento de los ojos en direcciones opuestas. Su finalidad es proyectar la imagen sobre ambas retinas y obtener una única imagen fusionada. Los movimientos de vergencia son de dos tipos: convergencia y divergencia. En el primer caso (convergencia) el movimiento de los ojos se dirige hacia la nariz y ocurre cuando el campo visual u objeto a explorar se acerca hacia el sujeto; en el segundo caso (divergencia) el movimiento de los ojos se produce hacia el exterior y aparece en el supuesto contrario. Ambos movimientos de vergencia llegan a alcanzar velocidades de unos 10 º / s. y su amplitud alcanza los 15 º de ángulo visual.

Se han descrito brevemente los movimientos oculares y su relación con el proceso de atención y los factores que influyen en el mismo, ya que su incorporación a los modelos del SVH (si bien compleja) puede resultar de gran interés en su aplicación a métricas de calidad perceptible.

En el siguiente capítulo se estudiarán algunas medidas de calidad cuyo modelado del sistema visual humano se basa en las propiedades analizadas previamente.

3

Evaluación de calidad en vídeo

En este capítulo se hará un breve repaso de algunos de los modelos existentes de medida de calidad de imagen y vídeo, comentando su estructura básica y los distintos problemas y ventajas que llevan asociados. En cualquier caso, no se trata de un análisis en profundidad de los diferentes modelos, sino una explicación breve con los detalles de mayor relevancia dentro de cada método particular. Se proporciona así un marco general en el que se circunscribe este proyecto. Para una compresión más completa de cada uno de los diferentes modelos que existen en la literatura se debe consultar la bibliografía. Previamente se indicarán las degradaciones y artefactos que introducen los algoritmos de compresión utilizados actualmente en la codificación digital de vídeo; detallando sus principales características y su influencia en la calidad percibida.

En primer lugar, se debe establecer una definición de calidad de imagen y vídeo. Es imprescindible comprender lo que significa este término para el observador ya que la forma de diseñar los modelos de medida se debe corresponder con lo que se entiende por calidad. Se define calidad como propiedad o conjunto de propiedades inherentes a una cosa. Este concepto, llevado a imágenes o vídeo, se entendería como la capacidad que una secuencia de vídeo tiene de representar el objeto original, es decir, la exactitud o parecido entre

Las imágenes de degradaciones incluidas en el apartado 3.1 han sido extraídas de [55]. Por su parte, las figuras correspondientes a los apartados 3.3 y 3.4 se han extraído de [65].

ambos. Dentro de los modelos de medida de calidad, en lugar de tener un vídeo y un objeto, la definición se extiende a dos secuencias de las cuales una es considerada la de referencia u original (hace las veces de objeto) y otra sobre la que se ha efectuado alguna operación de compresión, procesado, etc. En este caso, la medida de calidad de vídeo es una medida de semejanza entre la secuencia original y la distorsionada. Por tanto, para poder evaluar la calidad de un vídeo en todos los modelos que aparecen en este capítulo será necesario disponer de la secuencia original (FR). El desarrollo de modelos de medida de calidad en los que no exista imagen de referencia (NR), o sólo exista parcialmente (RR), es realmente complejo. Aún se están realizando los primeros estudios teóricos en este campo, con reducidas experiencias prácticas, por lo que se sitúa fuera del contexto de este proyecto, aunque el sistema visual sea capaz de realizar esta tarea de forma natural. Se han desarrollado algunas técnicas de medida NR y RR para imágenes, pero la investigación en vídeo se encuentra aún en una fase realmente temprana.

La definición de calidad que se ha dado corresponde con lo que se conoce como fidelidad, y presenta también algunos inconvenientes. En realidad, no está claro que la visibilidad del error esté relacionada con pérdida de calidad y, de hecho, algunas distorsiones que pueden ser claramente visibles no son, a pesar de ello, molestas al observador. Por ejemplo, si se considera la medida de calidad entre dos imágenes, una de las cuales es la multiplicación de los valores de luminancia de la otra imagen por un factor global, la diferencia visual entre ambas será obvia aunque el observador no aprecie tal diferencia como pérdida de calidad.

En segundo lugar, en cuanto a la clasificación de los modelos de medida, se van a diferenciar dos grupos principales: métricas de calidad de imagen subjetivas (utilizando observadores) y objetivas (mediante medidas matemáticas). Dentro de las objetivas se hará una nueva clasificación atendiendo únicamente a si se incorpora o no el SVH descrito previamente. Así, la clasificación resultante quedará como muestra el esquema de la siguiente página.

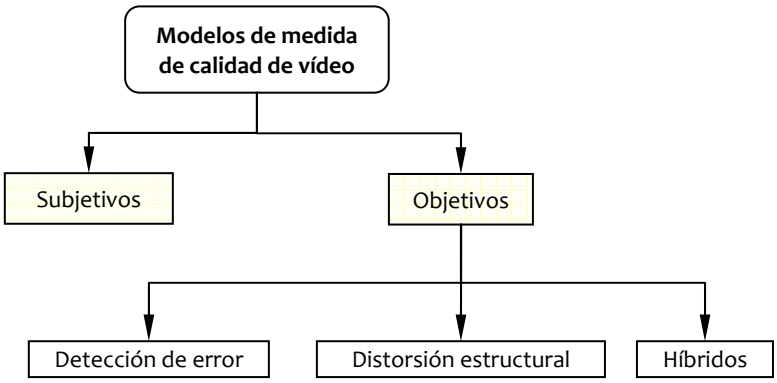


Fig. 3.1 Clasificación de los modelos de medida

Los modelos de *detección de error* obtienen una medida relativa a la diferencia de las imágenes entre sí pudiendo incorporar en mayor o menor medida las características propias del SVH. En un segundo grupo encontramos una nueva filosofía basada en la *distorsión estructural* existente entre las secuencias. Finalmente, existe un tercer grupo de modelos *híbridos* de medida de calidad, en el que se describen algunos métodos de medida que a pesar de no incluir un modelo de SVH tratan de obtener valores que se adecuen a la calidad que el observador percibe explotando otras cualidades relativas a la percepción. Existen algunos métodos de este último grupo para evaluación de calidad de imagen, pero en vídeo, los métodos híbridos aún no son más que conceptos teóricos que se están debatiendo en el VQEG, sin ninguna técnica eficaz.

Por último, se van a enumerar algunas de las características o propiedades de los modelos de calidad sobre las que se puede evaluar el rendimiento del método seleccionado. Entre ellas se pueden destacar las siguientes:

- *Velocidad:* es deseable que el valor de calidad que resulta de la utilización del método se obtenga de forma rápida. Esto adquiere un especial interés cuando se va a hacer uso de la medida de calidad para mejorar los procesos o algoritmos de compresión, cuantificación, etc.

- *Coste:* la carga computacional depende de la velocidad y de la complejidad de los algoritmos para la obtención de un resultado. Además, se deben considerar otros costes adicionales ya que para la validación o evaluación del método es necesario

llevar a cabo pruebas psicofísicas que tienen una serie de requisitos temporales y también en cuanto a número y características de los observadores.

- *Complejidad:* está estrechamente ligada con la velocidad y el coste. Lo ideal sería encontrar un método lo más sencillo posible que diese una medida de calidad igual a la percibida. En general, resultados más próximos a los obtenidos por un observador se consiguen al incorporar características propias del SVH, lo que forzosamente implica que el método sea complejo.

- *Portabilidad:* los resultados que proporciona el método no deben alterarse si se repiten las medidas en diferentes entornos o tiempos.

- *Precisión:* referida a cómo representa el resultado del método de medida la percepción de calidad que tendría el observador.

- *Robustez:* se pretende obtener resultados válidos sobre un amplio margen de variación de los parámetros asociados a la medida (tipo de imagen, tipo de distorsión, condiciones de visibilidad, etc.), es decir, se buscan métodos robustos.

- *Forma del resultado:* pueden ser valores numéricos (índices de calidad) o mapas de visibilidad del error. Según la aplicación a la que esté destinado un método de medida será conveniente una forma u otra y en general, lo ideal es que el método pueda proporcionar ambas salidas.

3.1 Distorsiones de vídeo y artefactos de codificación

El proceso de digitalización de vídeo utiliza técnicas que transforman un conjunto de píxeles a un dominio de frecuencia espacial (DCT), cuantificando valores, descartando eventualmente componentes de alta frecuencia, y haciendo uso de técnicas de predicción y compensación de movimiento. Esto genera ciertas distorsiones y artefactos de codificación [54], que pueden alterar la imagen original hasta niveles perceptibles, provocando las clásicas degradaciones que pueden verse en imágenes y vídeos con alta compresión.

Los algoritmos de compresión utilizados actualmente en la codificación digital de vídeo introducen varios tipos de degradaciones, que se pueden clasificar según sus características principales [55]. Esta clasificación es útil para poder comprender sus causas así como el impacto que tienen en la calidad percibida.

Por otro lado, al enviar el vídeo digital por algún medio de transmisión no fiable (como redes de paquetes) se suman adicionalmente degradaciones introducidas por las características propias del medio (por ejemplo, retardos o pérdidas de paquetes). La combinación de ambas degradaciones es la que finalmente percibe el usuario.

A continuación se presentan los numerosos artefactos de codificación producidos en la compresión digital de vídeos. Esta clasificación es de vital importancia a la hora de diseñar un algoritmo de evaluación subjetiva de calidad de vídeo. Es necesario identificar qué tipo de distorsiones afectan más a la calidad percibida, en qué punto un artefacto en concreto comienza a ser visible, y en general, todos aquellos aspectos que influyan en la percepción general de la calidad subjetiva.

No obstante, cuantificar los artefactos en función de su impacto visual no es una tarea sencilla. Debido a la complejidad del sistema visual humano, que como se ha visto, aún no ha sido modelado completamente, la distorsión percibida no es directamente proporcional al error absoluto, sino que depende tanto del tipo de distorsión presente como de las características espaciales y temporales de la secuencia de vídeo.

3.1.1 Efecto de bloques (blocking)

El efecto de bloques es, quizás, la más notoria de las degradaciones percibidas en vídeo digital. Tiene su origen en la codificación basada en bloques que segmenta la imagen en áreas pequeñas realizando una transformación de cada una de ellas de forma independiente. Visualmente, el efecto se observa como una discontinuidad entre los bloques adyacentes de una imagen. Se puede observar un ejemplo en la figura 3.2.

El umbral de cuantificación a partir del cual es percibido el efecto de bloque depende del tipo de imagen y del movimiento, por lo que no es posible definir un valor estándar e independiente de otros factores. Generalmente el efecto es menos percibido en imágenes con movimiento, o en áreas con una luminancia muy alta o muy baja.

Fig. 3.2 Ejemplo de efecto de bloques. Es más evidente en zonas de luminancia media y de textura uniforme.

Una de las razones principales para codificar los píxeles en bloques es aprovechar la alta correlación local entre los píxeles de una imagen. Sin embargo, al codificar cada bloque como unidad independiente, no se tiene en cuenta la posibilidad de que la correlación de los píxeles se pueda extender más allá de los bordes del bloque hasta los bloques adyacentes, lo que da lugar a discontinuidades en los bordes.

La detección y cuantificación del efecto de bloques en imágenes y vídeos es uno de los índices más comúnmente utilizado para estimar la calidad perceptual en medidas sin referencia (de tipo NR).

3.1.2 Efecto imagen de base (basis image)

El importante impacto visual del efecto de bloques se debe principalmente a la regularidad de tamaño y espaciado en las discontinuidades del borde de los bloques. Cada uno de estos bloques se representa mediante los coeficientes de su transformada. Sin embargo, en los casos en los que uno de los coeficientes de la DCT es muy superior al resto,

y al utilizar cuantificaciones altas, es posible que quede como resultado un único coeficiente, que se traduce, al decodificar, como uno de los 63 posibles patrones de imágenes base de la DCT. Al igual que en el efecto de bloques, el patrón regular de las imágenes base, así como su tamaño fijo, provoca que este efecto sea especialmente visible.

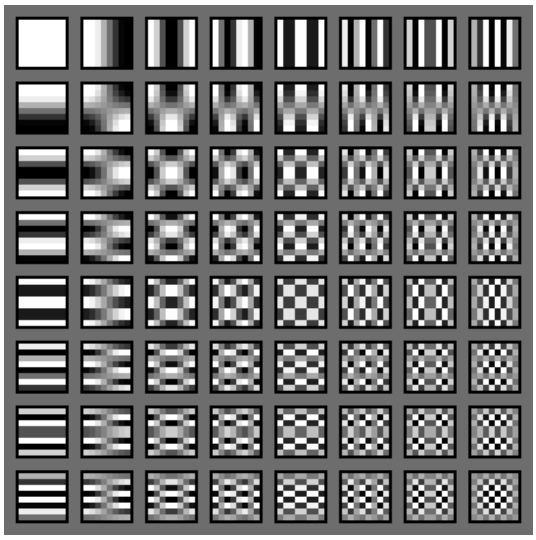


Fig. 3.3 Bases de la DCT 8x8 2D

Las características visuales de este efecto pueden además producir otros artefactos de codificación al considerar los bloques adyacentes. Un bloque que ha sufrido un efecto de imagen base, seguramente no se corresponderá con sus bloques adyacentes, acentuando el efecto de bloques, y el efecto de mosaico, que se detalla más adelante. En la figura 3.4 podemos observar un ejemplo de efecto imagen base.

Fig. 3.4 Ejemplo de efecto de imagen base, extraído del fondo de la secuencia *Table-Tennis*

3.1.3 Desenfoque o falta de definición (blurring)

La falta de definición se manifiesta como una pérdida de detalle espacial en la imagen, visible principalmente en áreas de actividad espacial alta o moderada como texturas detalladas o bordes de objetos en la escena. Si bien puede producirse por imágenes tomadas fuera de foco, también puede ser un efecto introducido por el proceso de digitalización. En este caso, se da al suprimir los coeficientes DCT de mayor orden, que son los que aportan los detalles finos dentro de sus bloques. Esta degradación también puede hacer más visible el efecto de bloques y de mosaico.

En macrobloques de tipo P (predicted), el desenfoque es principalmente consecuencia del uso de un macrobloque referencia con pérdida de detalle espacial. Sin embargo, también puede ocurrir en macrobloques de tipo B (bidirectional), donde la interpolación entre las predicciones hacia adelante y hacia atrás puede promediar los contenidos en la predicción final del macrobloque.

3.1.4 Desplazamiento de color (color bleeding)

Como se ha visto en el apartado anterior, el desenfoque en la información de luminancia resulta en pérdida de detalle espacial. El correspondiente efecto para la información de crominancia resulta en una contaminación de color en áreas con un nivel alto de contraste en crominancia.

Al igual que el caso anterior, este efecto se debe a la supresión de los coeficientes de alta frecuencia de los componentes de crominancia. Dado que la información de color es sub-muestreada, el efecto no se limita al bloque de 8x8 píxeles, sino que se propaga a todo el macro-bloque.

En la figura 3.5 se muestra un fotograma de la secuencia de prueba *Table-Tennis*, codificada a 0,6 mbps. El desenfoque de las componentes de crominancia es más evidente a lo largo de la parte inferior del brazo, así como en la pala. Se puede observar como el efecto del desplazamiento de color afecta al macrobloque completo.

Fig. 3.5 Fotograma de la secuencia *Table-Tennis* en la que se aprecia desplazamiento de color

3.1.5 Efecto escalera (staircase effect)

Las imágenes base de la DCT no están adaptadas a la representación de bordes diagonales. Por lo tanto, son necesarias imágenes base de muy alta frecuencia para representar adecuadamente bordes o características diagonales en las imágenes. Tras una cuantificación agresiva, el truncamiento de las contribuciones hechas por los coeficientes de mayor orden impide contrarrestar las contribuciones de las imágenes base de menor orden en el bloque reconstruido. Esto provoca que un borde diagonal orientado hacia la horizontal se reconstruya con un borde horizontal, y viceversa para los bordes orientados hacia la vertical. El efecto escalera se da al representar un borde diagonal como una concatenación de bloques horizontales y/o verticales.

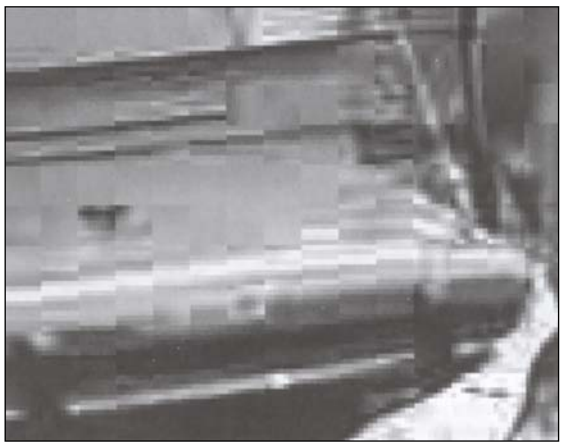


Fig. 3.6 Ejemplo de efecto escalera

3.1.6 Ringing

La representación de un bloque puede considerarse una combinación ponderada de cada imagen base DCT, de forma que las contribuciones de una imagen base se acentúan o se contrarrestan con las contribuciones de otras imágenes base. Por lo tanto, la cuantificación de un coeficiente individual resulta en un error en la contribución de la imagen base correspondiente. Al tener las imágenes de alta frecuencia una importancia mayor en la representación de bordes, la reconstrucción de bloques cuantificados incluirá irregularidades de alta frecuencia.

El efecto de ringing es más evidente en bordes de alto contraste presentes en texturas lisas; aparece un ruido en alta frecuencia desde el borde hasta el límite del bloque que lo contiene. Cuanto mayor es el contraste del borde, mayor es el nivel de los picos. En las figuras 3.7 y 3.8 se muestran ejemplos de ringing. El efecto se hace más evidente en el borde de la mesa de ping-pong y en la parte inferior del brazo del jugador. La figura 3.8 muestra una sección de un cuadro de tipo I de la secuencia *Claire* codificada a 1.0 Mbps. Aunque está codificada a una tasa alta, aún aparece un ligero efecto de ringing a lo largo del borde del brazo de la presentadora (de alto contraste).

Fig. 3.7 Ejemplo de efecto ringing en un fotograma de *Table-Tennis*

Fig. 3.8 Ejemplo de efecto ringing en un fotograma de *Claire*

3.1.7 Patrones de mosaico (mosaic patterns)

La consecuencia general del efecto de patrones de mosaico es la aparente discordancia entre todos, o parte, de bloques adyacentes en una imagen. Tiene un efecto similar al producido al usar piezas cuadradas que no corresponden en un mosaico. El patrón de mosaico generalmente coincide con el efecto de bloques. Se produce a causa de un bloque con un cierto contorno o textura muy distinta a los bloques vecinos.

Fig. 3.9 Ejemplo de efecto de patrones de mosaico. Más evidente alrededor de la cara y cerca de los bordes horizontales de la camioneta.

3.1.8 Falso contorno (false contouring)

El falso contorno resulta a menudo de la cuantificación directa de los valores de los píxeles. Ocurre en zonas de textura lisa que contienen una transición gradual. La operación de cuantificación restringe los valores de los píxeles a un conjunto reducido del rango original, lo que resulta en una serie de gradaciones escalonadas al reconstruir la imagen.

El falso contorno puede ocurrir también a causa de las transformaciones basadas en bloque. Es consecuencia de una cuantificación inadecuada del coeficiente de continua y de los coeficientes de baja frecuencia. En la reconstrucción se observa el efecto como transiciones escalonadas en zonas donde originariamente existían transiciones suaves. Se muestra un ejemplo en la figura 3.10, que contiene una sección de un cuadro I de la secuencia Claire. El fotograma original contenía una reducción gradual de la luminancia desde la cabeza de la presentadora.

Fig. 3.10 Ejemplo de falso contorno

3.1.9 Falsos bordes (false edges)

Este efecto se presenta como consecuencia de transportar el efecto de bloques hacia cuadros predictivos, con compensación de movimiento. Si se produce el efecto de bloques en una imagen tomada como referencia para próximos cuadros, estas discontinuidades producidas entre los bloques, pueden ser convertidas en falsos bordes dentro de bloques predictivos, debido a la estimación de movimiento. El origen de este efecto se esquematiza en la figura 3.11. Se puede ver como los bordes de los bloques se trasladan en cuadros predictivos (B, P) a regiones que quedan dentro de los bloques, produciendo el efecto de falsos bordes.

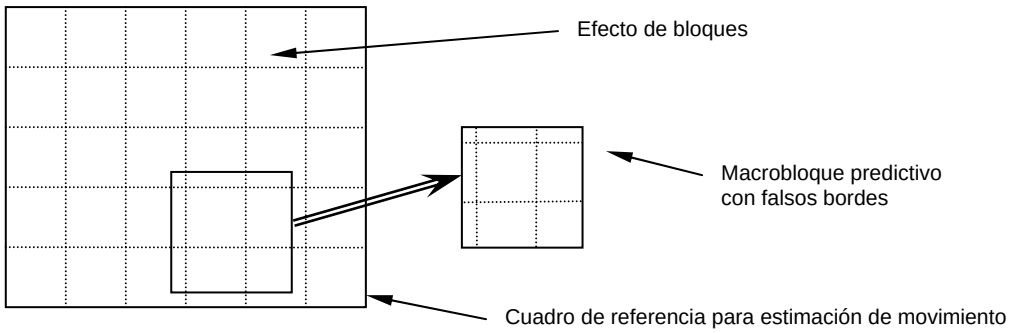


Fig. 3.11 Ejemplo de propagación del efecto de bloques en falsos bordes

3.1.10 Errores de compensación de movimiento (MC mismatch)

La estimación de movimiento de un macrobloque se realiza en el codificador, comparando el macrobloque de un fotograma con todas las posibles secciones de igual tamaño (dentro de cierto rango espacial) de la(s) imagen(es) adyacente(s). La comparación se realiza generalmente buscando el mínimo valor de MSE entre el macrobloque y la sección evaluada. Este procedimiento (simple, pero de alto coste computacional), se basa en la hipótesis que todos los píxeles del macrobloque tendrán un mismo desplazamiento, es decir, que corresponden a la misma figura en la imagen. Esto no es así cuando el macrobloque contiene partes de diferentes figuras, o cuando contiene el borde entre una figura y un fondo fijo. En estos casos, la estimación de movimiento no será adecuada para

una o quizás para ninguna de las figuras dentro del macrobloque. En estos casos, al reconstruir un macrobloque basado en una mala estimación de movimiento, se puede ver un efecto de bloque, con componentes de alta frecuencia espacial, que generalmente se aprecia sobre los bordes de figuras en movimiento.

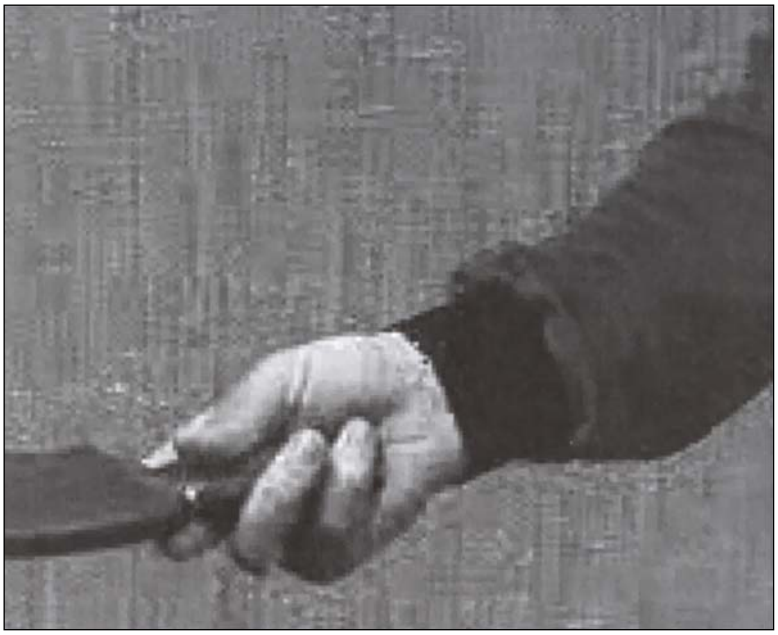


Fig. 3.12 Ruido de alta frecuencia a causa de errores de CM alrededor de los objetos en movimiento

3.1.11 Efecto mosquito (mosquito effect)

El efecto mosquito es un artefacto temporal que se observa principalmente en texturas lisas como fluctuaciones en los niveles de luminancia o crominancia alrededor de bordes de alto contraste u objetos en movimiento. Este efecto está relacionado con las distorsiones en alta frecuencia introducidas por el efecto de ringing y los errores de predicción a causa de fallos en la compensación de movimiento. Generalmente la visibilidad de las fluctuaciones es menor si el efecto es causado por ringing. No obstante, con independencia del origen, normalmente es consecuencia de una codificación diferente de la misma área de una escena en fotogramas consecutivos. La codificación puede diferir en el tipo de predicción (*forward*, *backward*, *bidireccional*, o *skipped*), en el nivel de cuantificación, en la compensación de movimiento, o en una combinación de estos factores.

3.1.12 Fluctuaciones en áreas estacionarias

Fluctuaciones similares a las del efecto mosquito pueden verse también en áreas sin movimiento, pero con gran contenido de alta frecuencia espacial (por ejemplo fondos con detalles pequeños, texturas complejas, etc.) Como en el caso anterior, las fluctuaciones son consecuencia de los diferentes tipos de predicciones y niveles de cuantificación utilizados entre cuadros. Estos efectos pueden ser enmascarados en áreas con movimiento, y por lo tanto sólo se perciben en imágenes estáticas.

3.1.13 Errores de crominancia (chrominance mismatch)

Como se mencionó al explicar los errores de compensación de movimiento, la elección de los vectores de movimiento se basa en estimar el desplazamiento en la escena con el fin de minimizar los errores MSE de la luminancia. Generalmente, la crominancia no se tiene en cuenta en esta estimación, aunque posteriormente los valores estimados del movimiento se utilizan para las tres componentes de vídeo. Esto puede provocar que la estimación no se adecue a la realidad, teniendo como consecuencia la aparición de macrobloques cuyos colores no corresponden con los originales.

3.2. Medidas de distorsión comparativa

Como ya se comentó, las métricas de calidad objetiva principalmente empleadas durante muchos años (y que aún se siguen empleando) han sido simples medidas matemáticas. En efecto, al modelar una imagen o un vídeo cuya calidad desea ser evaluada como suma de una señal perfecta más una señal de error, se puede asumir que la pérdida de calidad está directamente relacionada con la potencia de la señal de error. Por lo tanto, una forma natural de evaluar la calidad de la imagen es cuantificar el error entre la señal distorsionada y la original. A continuación se presentan las métricas de este tipo más comunes en la mayoría de las aplicaciones de tratamiento de imagen.

MSE – Mean Squared Error

El error cuadrático medio es la implementación más simple y más extendida de las medidas basadas en error.

$$MSE = \frac{\sum_{i=0}^{N}\sum_{j=0}^{M}\left(x(i,j)-d(i,j)\right)^2}{NM} = \frac{(\mathbf{x}-\mathbf{d})(\mathbf{x}-\mathbf{d})^T}{NM}$$

siendo: N, M la dimensión vertical y horizontal de la imagen en píxeles

$x(i,j)$, $d(i,j)$ imagen original y distorsionada respectivamente en notación matricial

$\mathbf{x}$ y $\mathbf{d}$ imagen original y distorsionada respectivamente en notación lexicográfica

RMSE – Root Mean Squared Error

Raíz del error cuadrático medio. $RMSE = \sqrt{MSE}$

NMSE – Normalized Mean Squared Error

Error cuadrático medio normalizado.

$$NMSE = \frac{\sum_{i=0}^{N}\sum_{j=0}^{M}\left(x(i,j)-d(i,j)\right)^2}{\sum_{i=0}^{N}\sum_{j=0}^{M}\left(x(i,j)\right)^2} = \frac{(\mathbf{x}-\mathbf{d})(\mathbf{x}-\mathbf{d})^T}{\mathbf{x}\mathbf{x}^T}$$

SNR – Signal to Noise Ratio

Relación señal a ruido. $SNR(dB) = -10\cdot\log_{10}(NMSE)$

PSNR – Peak Signal to Noise Ratio

Relación señal a ruido de pico. $PSNR(dB) = 10\cdot\log_{10}\frac{L^2}{MSE}$

donde L es el rango dinámico de los valores de los píxeles. Para una señal monocromática de 8 bits/píxel, L es igual a 255.

SER – Signal to Error Ratio

Relación señal a error. $SER = 20\cdot\log_{10}\left(\frac{x(i,j)_{max}}{RMSE}\right)$

STD – Standard Deviation

Desviación típica.

$$STD = \frac{1}{NM}\sum_{i=0}^{N}\sum_{j=0}^{M}\left(\left(x(i,j)-d(i,j)\right)-\mu\right)^2 = \frac{1}{NM}\left(\left(\mathbf{x}-\mathbf{d}\right)-\mu\right)\left(\left(\mathbf{x}-\mathbf{d}\right)-\mu\right)^T$$

donde $\mu = \dfrac{\sum_{i=0}^{N}\sum_{j=0}^{M}\left(x(i,j)-d(i,j)\right)}{NM}$

Tales índices forman un conjunto de parámetros de calidad objetiva conocidos como *Medidas de Distorsión Comparativa*, y explican el hecho de que grandes valores de error corresponden a una mala calidad de imagen. De ellas, las más usadas han sido el error cuadrático medio (*MSE*) y la relación señal a ruido de pico (*PSNR*) ya que son simples de calcular, tienen significados físicos claros y son fáciles de tratar matemáticamente. Sin embargo, como se explicó anteriormente, al no correlacionar bien con la calidad percibida han sufrido numerosas críticas [1-4]. El problema principal que presentan estas medidas es que sólo son eficientes cuando los errores se comportan como ruido adicional no correlacionado con la señal. Una mejor estimación de la importancia del error para el ojo humano se consigue al aplicar las medidas de distorsión comparativa a imágenes previamente filtradas con filtros de modelado del Sistema Visual Humano, como se analizará en el apartado siguiente.

3.3. Medidas basadas en detección de error

La mayoría de las medidas de calidad basadas en el SVH también comparten un paradigma basado en la detección de error, motivado por la investigación en visión psicofísica y psicológica. El principio básico es considerar la secuencia distorsionada como suma de una señal de referencia de calidad perfecta y una señal de error. Los algoritmos de evaluación de calidad de vídeo, por tanto, han de determinar la potencia de la señal de error y analizar en qué medida afecta a la percepción humana, según las características del SVH.

La figura 3.13 muestra el esquema general de un algoritmo de medida de calidad basado en detección de error y modelado según el HVS. Casi todos los algoritmos que modelan el SVH pueden explicarse con este esquema, aunque pueden diferir en detalles específicos.

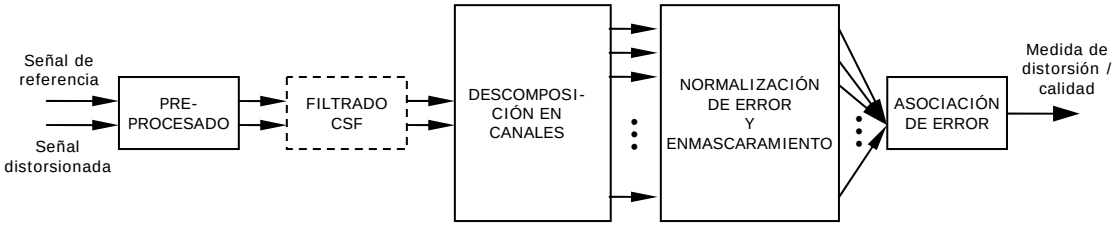


Fig. 3.13 Esquema de un sistema de medida de calidad basado en detección de error. La característica CSF puede ser implementada como “filtrado CSF” o como “normalización de error”

• *Preprocesado*

La etapa de pre-procesado puede llevar a cabo las siguientes operaciones: alineado espacial, alineado temporal, transformaciones de espacio de color, calibración para dispositivos de visualización, filtrado PSF y adaptación de luz. En primer lugar, es necesario alinear correctamente las señales de referencia y distorsionada. La señal distorsionada puede desalinearse con respecto a la original, de forma local o de forma general por múltiples razones durante los procesos de compresión, procesado y transmisión. Por ello, es necesario establecer una correspondencia punto a punto entre ellas. Tras ello, normalmente se prefiere transformar las señales a un espacio de color que se adapte mejor al SVH. Posteriormente, las medidas de evaluación de la calidad pueden necesitar convertir los valores digitales de los píxeles, almacenados en la memoria del equipo, a valores de luminancia de los píxeles de un dispositivo de visualización a través de transformaciones no lineales de cada punto. Después, se puede aplicar un filtro paso bajo para simular la PSF de la óptica del ojo. Por último, los vídeos de referencia y distorsionado necesitan ser convertidos a su correspondiente estímulo de contraste para simular la adaptación a la luz. Muchos modelos trabajan con contraste de banda limitada para escenas naturales complejas, ligado con la descomposición de canal. En ese caso, el cálculo del contraste se implementa en el sistema más adelante, durante o después del proceso de descomposición en canales.

• *Filtrado CSF*

La función de sensibilidad al contraste puede ser implementada antes de la descomposición de canal usando filtros lineales que aproximen la respuesta en frecuencia de la CSF. Sin embargo, algunas medidas optan por implementar el filtrado como factores de ponderación de cada uno de los canales después de la descomposición en canales. Es decir, la salida de cada canal o banda se pondera en función de la sensibilidad a la correspondiente frecuencia.

• *Descomposición en canales*

Las medidas de calidad normalmente modelan la descomposición en canales restringiéndose a las limitaciones de la aplicación y a la carga computacional. Los canales sirven para separar los estímulos visuales en diferentes subbandas espaciales y temporales. Mientras que algunos algoritmos de evaluación de la calidad implementan sofisticados sistemas de descomposición en canales, algunas transformaciones simples como la trasformada wavelet, o incluso la transformada discreta del coseno (DCT) se han utilizado en los primeros momentos en la literatura debido a su utilidad en ciertas aplicaciones, más que por su precisión, como por ejemplo en el modelado de las neuronas corticales.

Aunque las zonas corticales encargadas de la visión se representan bien con funciones 2D de Gabor, esta descomposición es muy compleja de calcular y carece de algunas de las características matemáticas deseables para una correcta implementación, como invertibilidad, reconstrucción por adición, etc. Watson implementó la *cortex transform* (transformación de la corteza) [19] para modelar la descomposición en canales selectiva en frecuencia y orientación, similar a las funciones 2D de Watson pero con una implementación más viable. Los modelos de descomposición de canal usados por Watson, Daly [13], Lubin [20,21] y Teo y Hegger [22,23] tratan de modelar el SVH lo más fielmente posible sin incurrir en dificultades prohibitivas en su implementación. Las configuraciones de subbandas para algunos de estos modelos se proporcionan en la figura 3.14.

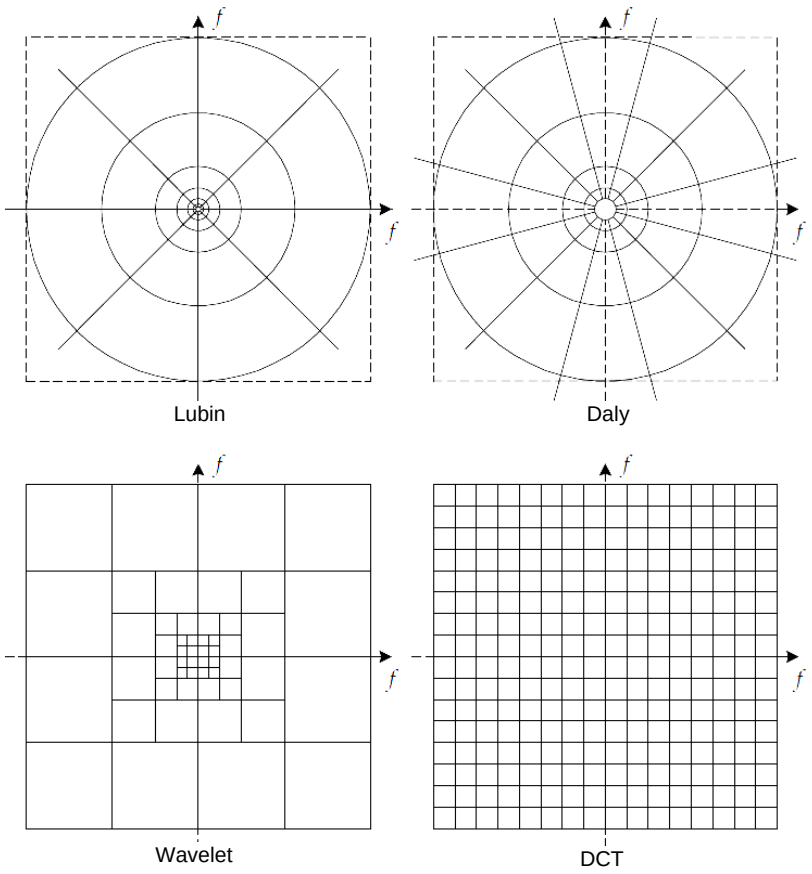
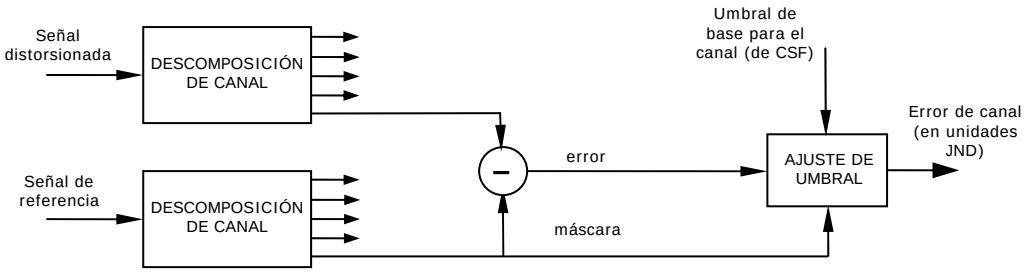


Fig. 3.14 Descomposición en frecuencia de varios modelos

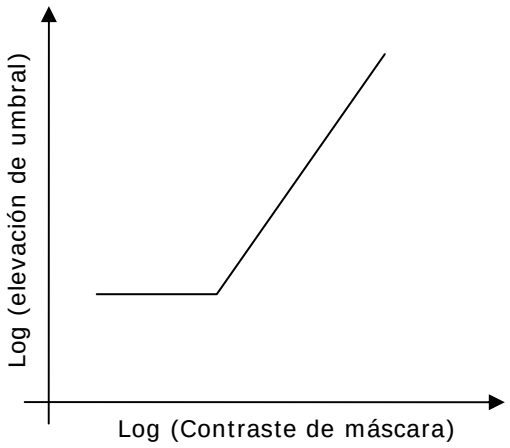
• *Normalización de error y enmascaramiento*

Estas características se incluyen normalmente en cada canal. La mayoría de los modelos implementan el enmascaramiento como un mecanismo de control de ganancia que pondera la señal de error en cada canal mediante un umbral de visibilidad variante en el espacio para cada uno de los canales [24]. El ajuste del umbral de visibilidad en un punto se calcula basándose en la energía de la señal de referencia (o ambas, señal de referencia y distorsionada) en un entorno reducido de dicho punto, así como la sensibilidad del SVH para ese canal en ausencia de los efectos de enmascaramiento (conocida como sensibilidad base). La figura 3.15 a. muestra como se implementa normalmente el enmascaramiento en un canal. Para cada uno de los canales el umbral de error base (el contraste mínimo visible del error) se eleva para tener en cuenta la presencia de la señal enmascarante. La relación entre la elevación de umbral y el contraste de la señal referencia (o distorsionada) en cada

canal es la mostrada en la figura 3.15 b. El umbral de visibilidad se utiliza luego para normalizar la señal de error. Esta normalización típicamente convierte el error en unidades JND, es decir, diferencia apreciable, donde un JND de 1.0 indica que la distorsión en ese punto, en ese canal está justo en el umbral de visibilidad. Algunos métodos implementan el enmascaramiento como representación de la saturación en la respuesta al contraste. La figura 3.16 muestra un conjunto de curvas cada una de las cuales representa las características de saturación de las neuronas del SVH. Las medidas pueden modelar el enmascaramiento con una o más de estas curvas.



a. Implementación del enmascaramiento



b. Modelado umbral de visibilidad

Fig. 3.15 (a) Implementación del efecto de enmascaramiento por canal basado en modelos HVS
(b) Modelado umbral de visibilidad (simplificado); elevación de umbral en función de contraste de máscara

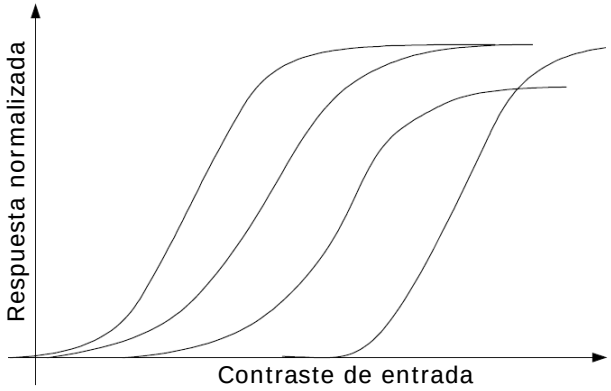


Fig. 3.16 Efectos de saturación no lineales en la respuesta

• *Combinación de error*

Es el proceso de asociar las señales de error de diferentes canales en una única interpretación de calidad o distorsión. En la mayoría de los métodos de evaluación de calidad, la combinación toma la siguiente forma:

$$E = \left(\sum_{l} \sum_{k} \left| e_{l,k} \right|^{\beta} \right)^{1/\beta}$$

donde $e_{l,k}$ es el error enmascarado y normalizado del coeficiente k-ésimo en el canal l-ésimo, y β es una constante con un valor normalmente entre 1 y 4. Este método es conocido como asociación de error Minkowski y puede realizarse espacialmente (índice k) y frecuencialmente (índice l), o viceversa, con algunas no-linealidades entre ellas, o posiblemente con diferentes exponentes β. También puede usarse un mapa espacial indicando la importancia relativa de las diferentes regiones para proporcionar una ponderación espacial a los diferentes $e_{l,k}$ [25,26].

3.3.1. Evaluación de calidad de imagen por detección de error

La mayoría de los esfuerzos de los investigadores se han centrado en el problema de evaluación de la calidad de imágenes, y sólo recientemente la evaluación de la calidad del vídeo ha recibido más atención. Las medidas de calidad de vídeo actuales usan modelos

SVH similares a los que se usan en muchos algoritmos de medida de la calidad de imagen, con extensiones apropiadas para incorporar los aspectos temporales del SVH. Por ese motivo, a continuación se presentan algunas métricas de calidad de imagen que están basadas en el paradigma de detección de error y servirán para posteriormente entender mejor los algoritmos de evaluación de vídeo. Es importante recordar que no se trata de un análisis en profundidad y detallado de los diferentes modelos, sino una explicación breve con los aspectos de mayor relevancia dentro de cada método particular.

El predictor de diferencias visibles (VDP, *Visible Differences Predictor*) de Daly [27] tiene como objetivo calcular un mapa de probabilidad de detección entre la señal de referencia y la distorsionada. El valor en cada punto del mapa es la probabilidad de que un observador humano note diferencia entre la señal original y la distorsionada en ese punto. Las dos imágenes (expresadas en valores de luminancia) son pasadas a través de una serie de procesos: operaciones puntuales no lineales, filtrado CSF, descomposición de canal, cálculo del contraste, modelado del efecto de enmascaramiento, y cálculo de la probabilidad de detección. Para la descomposición de canal se usa una transformada cortex modificada [19], que transforma la imagen en cinco niveles espaciales seguidos de seis niveles de orientación, resultando un total de 31 canales independientes (incluyendo la banda base). Para cada canal, se calcula un mapa de umbral de elevación para su contraste. Una función psicométrica se usa para convertir las señales de error (ponderadas por los umbrales de elevación) en un mapa de probabilidad de detección para cada canal. Posteriormente se realiza una combinación de todos los canales para obtener un mapa general de detección.

El algoritmo de Lubin [20,21] también pretende estimar la probabilidad de detección de las diferencias entre las versiones original y distorsionada. Se aplica un desenfoque para modelar la PSF de la óptica del ojo. Las señales son posteriormente remuestreadas para simular el muestreo de los fotorreceptores de la retina. Una pirámide laplaciana [28] se usa para descomponer las imágenes en siete niveles de resolución (cada uno es la mitad de la inmediatamente superior), para después realizar un cálculo de contraste limitado en banda. Un conjunto de filtros orientables de Freeman y Adelson [29] son entonces aplicados para la selección de orientación en las cuatro direcciones. La CSF se modela normalizando la salida de cada canal selectivo en frecuencia mediante la sensibilidad base de cada canal. El enmascaramiento se implementa a través de una función sigmoide no lineal, después de la

cual los errores son convolucionados con elementos estructurantes discoidales a cada nivel. Por último se realiza una combinación en un mapa de distorsión a través de la frecuencia. Una última combinación puede aplicarse para obtener un único valor que represente la imagen completa.

La métrica de Teo y Heeger [22,23] usa modelado PSF, enmascaramiento de luminancia, descomposición de canal, y normalización de contraste. El proceso de descomposición en canales usa filtros orientables en cuadratura [30] con seis niveles de orientación y cuatro resoluciones espaciales. Se implementa un mecanismo de detección basado en el error cuadrático. El enmascaramiento se modela mediante normalización de contraste y saturación de respuesta. La normalización de contraste es diferente del método de Lubin o Daly. Éstas toman las salidas de todos los canales en todas las orientaciones a una determinada resolución para llevar a cabo la normalización. Por el contrario, este modelo no asume que los canales a la misma resolución sean independientes. Sólo los canales a distinta resolución son considerados independientes. La salida de la descomposición de canal después de la normalización de contraste se descompone en cuatro mediante las respuestas mostradas en la figura 3.16, con los parámetros óptimos para adecuase a los experimentos psicovisuales.

La medida de la DCT de Watson [31] se basa en una transformada DCT 8x8 comúnmente usada en compresión de imagen y vídeo. Al contrario que los modelos anteriores, este método divide el espectro en 64 subbandas uniformes (8 en cada dimensión cartesiana). Después de la DCT basada en bloques y de calcular los contrastes de las subbandas, se halla un umbral de visibilidad para cada coeficiente de subbanda en cada bloque usando la sensibilidad base de dicha subbanda. La sensibilidad base se obtiene empíricamente. Los umbrales son corregidos para luminancia y enmascaramiento de texturas. El error en cada subbanda es ponderado por el correspondiente umbral de visibilidad y combinado mediante la asociación Minkowski espacial. Posteriormente se realiza una combinación entre subbandas usando la fórmula de Minkowski con un coeficiente diferente.

El codificador perceptual de imagen de Safranek-Johnston [32] incorpora una medida de calidad usando una estrategia similar a la DCT de Watson. La descomposición de canal

usa un filtro de espejo cuadrático generalizado (GQMF, *Generalized Quadrature Mirror Filter*) [33] para el análisis y síntesis. Esta transformada divide el espectro en 16 subbandas uniformes (cuatro en cada dimensión cartesiana). Los métodos de enmascaramiento y combinación son similares a los usados en la medida de Watson.

Bradley [34] documentó un predictor de diferencias visibles mediante wavelet (WVDP), que es una simplificación del VDP de Daly descrito anteriormente. Usa la derivada de Watson de umbral de detección de ruido de cuantificación para unas transformadas wavelets 9/7 biortogonales [35] y las combina con un umbral de elevación y un esquema de detección de probabilidad psicométrico similar al de Daly. Otra medida basada en wavelet ha sido propuesta por Lai y Kuo [36]. Su medida se basa en la wavelet Haar y su modelo de enmascaramiento puede considerar interacciones de canal así como efectos de supraumbrales.

Las medidas de calidad propuestas anteriormente son medidas de valores escalares. Damera-Venkata y otros propusieron una medida para cuantificar el rendimiento en los sistemas de restauración de imágenes, en los cuales la degradación se modela como una distorsión lineal en frecuencia más ruido aditivo [37]. Dos métricas complementarias fueron desarrolladas para cuantificar estas distorsiones de forma separada. Observaban que si el ruido aditivo no estaba correlado con la imagen de referencia, entonces una medida de error de un sistema basado en SVH correlacionaría bien con la evaluación subjetiva. Usando un algoritmo de restauración adaptativo espacialmente, aislaban los efectos del ruido y la distorsión frecuencial lineal. El ruido se cuantifica usando una medida basada en SVH multicanal, tras lo cual una medida de distorsión cuantifica la distorsión espectral entre la referencia y la imagen del modelo de restauración.

Algunos investigadores han tratado de medir la calidad de imagen usando modelos de un solo canal con el modelo del efecto de enmascaramiento centrado en ciertos tipos de distorsión, como los artefactos (blocking). En [38] y [39], Karunasekera y Kingsbury propusieron una medida de calidad para artefactos de blocking. En primer lugar se realiza una detección de bordes en la imagen distorsionada. Con la imagen de referencia se calcula un mapa de actividad en las proximidades de los bordes, y una imagen de enmascarado de actividad de bordes se computa de forma que se resta peso a los bordes que ocurren en

zonas de alta actividad. Esta imagen de enmascarado de actividad de bordes se ajusta para el enmascaramiento de luminancia. Una transformación no lineal se aplica antes de la combinación. Los parámetros para el modelo se obtienen de los experimentos que miden la sensibilidad de observadores humanos a los artefactos de borde embebidos en patrones de test de banda estrecha.

Chou y Li [40] definieron la relación señal a ruido perceptible de pico (PSPNR, *Peak Signal to Perceptible Noise Ratio*) que es una medida de un único canal. Modelaron el enmascaramiento de luminancia y de actividad para obtener un perfil JND. La PSPNR tiene la misma definición que la PSNR, exceptuando que la expresión MSE se ajusta para el perfil JND.

Otra medida de canal único es la escala de calidad de imagen (PQS) de Miyahara [41], en la que un número de características que pueden capturar varias distorsiones se combinan en un único valor.

3.3.2. Evaluación de la calidad de vídeo por detección de error

Una forma obvia de implementar una métrica de calidad de vídeo es aplicar un algoritmo de medida de imagen fija fotograma a fotograma. Sin embargo, una aproximación más sofisticada modelará los aspectos temporales del SVH. Se han propuesto varios algoritmos para extender las características del SVH a las dimensiones de tiempo y movimiento. A continuación se proporciona una breve explicación de los mismos.

En [42], Tan y otros implementaron una métrica de distorsión de vídeo usando una medida de calidad de imagen seguido de un "emulador cognitivo" que modela los efectos temporales como el suavizado y enmascaramiento temporal de de la medida de calidad del fotograma. Se empleó *tracking asimétrico*, que modela el fenómeno de la tendencia de los humanos a percibir una transición de calidad de buena a mala de forma más fácil que una transición de calidad de pobre a buena.

Van den Branden Lambretch y otros extendieron el modelo SVH a tres dimensiones modelando la dimensión temporal de la CSF, y generando dos flujos visuales centrados en diferentes aspectos temporales del estímulo de la salida de cada canal espacial [43,44,45]. Los dos flujos modelan los mecanismos transitorios y permanentes del SVH. Su propuesta de métrica de calidad de imagen en movimiento (MPQM *Moving Picture Quality Metric*) consiste en una descomposición de canal en cuatro escalas, cuatro orientaciones y dos flujos temporales. Los canales de salida resultantes son restados para crear la señal de error. El enmascaramiento se implementa mediante normalización de los canales de error por los umbrales de visibilidad dependientes del estímulo (similares a los usados en las métricas de calidad de imágenes fijas).

En [3], Winkler presentó una medida de calidad para vídeo en color. El algoritmo usa una transformación del espacio de color y aplica la métrica en cada canal de color transformado. Se generan dos flujos temporales usando filtros IIR, con descomposición espacial en cinco niveles de subbandas y cuatro orientaciones. Los canales son ponderados por la correspondiente CSF, y el enmascaramiento se implementa basado en el modelo excitativo-inhibidor propuesto por Watson y Solomon [24].

La medida de calidad de vídeo digital (DVQ, Digital Video Quality) de Watson opera en el dominio de la DCT y, por ello, es más atractiva desde un punto de vista del coste computacional [46,47], ya que la DCT es muy eficiente en su implementación y la mayoría de los estándares de codificación de vídeo se basan en ella. Se propuso un modelo de umbral de visibilidad tridimensional para los canales espaciotemporales de la DCT. El algoritmo DVQ en primer lugar calcula la DCT de las señales de referencia y distorsionada, respectivamente. Posteriormente computa el contraste local, aplica el filtrado CSF temporal, y convierte los resultados a unidades JND normalizándolos con los umbrales de visibilidad, tras lo cual la señal de error se computa. Finalmente, el enmascaramiento y la combinación se aplican a las señales de error. DVQ implementa una transformación de color antes de aplicar la métrica a cada una de las dimensiones de crominancia.

Otra medida que modela los aspectos visuales del SVH se presentó por Tan y Ghanbari [48]. Trata de evaluar la calidad del vídeo MPEG y combina un modelo

perceptual típico basado en detección de error con un modelo de medida del efecto de bloques. El modelo perceptual consiste en una corrección gamma, operación puntual no lineal, cálculo del contraste, filtrado CSF espacial, filtrado temporal, descomposición frecuencial en dos canales (diagonal y horizontal/vertical), respuesta no lineal al contraste, promediado de error, enmascaramiento, asociación, promediado temporal y enmascaramiento de movimiento. El detector de efecto de bloques se basa en un análisis armónico de la señal de bordes de bloques, combinado con un modelo de enmascaramiento visual. El valor final de calidad es el modelo perceptual o la puntuación del detector de efecto de bloque.

En [49], Yu y otros propusieron una medida de calidad de vídeo basada en la extensión de la medida de distorsión perceptual de Winkler [3] hacia una medida de la distorsión perceptual del efecto de bloques. Los parámetros para los modelos se obtienen minimizando el error en las predicciones de calidad para secuencias de vídeo obtenidas de una base de datos de pruebas subjetivas del VQEG. Esto es en contraste a la mayoría de los métodos que obtienen los parámetros para ajustar los umbrales de los experimentos psicovisuales con patrones simples. Se ubica específicamente el artefacto de blocking mediante combinación espacial sobre aquellas áreas en las que los efectos del blocking son dominantes.

Existen numerosos aspectos en la implementación que han de ser considerados antes de desarrollar un sistema de medida de calidad de vídeo práctico. Un factor importante que afecta a la viabilidad de una métrica de calidad es su coste computacional. Mientras que métodos de medida complejos pueden modelar el SVH de forma más precisa, su complejidad computacional puede ser prohibitiva para numerosas plataformas, especialmente para medidas en tiempo real o en vídeo de alta resolución. Las necesidades de memoria son otro factor importante. Por ejemplo, para implementar el filtrado temporal, se puede necesitar una gran cantidad de memoria para almacenar un gran número de fotogramas, lo que encarece los desarrollos finales.

3.3.3. Limitaciones

El principio fundamental de los algoritmos basados en la detección de error es predecir la calidad perceptual cuantificando los errores perceptibles. Esto se consigue simulando la calidad perceptual relacionada con los componentes visuales del sistema visual humano. Sin embargo, el SVH es extremadamente complicado, un sistema altamente no lineal y el conocimiento actual del mismo es muy limitado. Hasta dónde pueden llegar los algoritmos basados en la detección de error es una cuestión que puede necesitar años para ser contestada.

La mayoría de las medidas de error realizan, de forma explícita o implícita, una serie de suposiciones. La siguiente es una lista incompleta (un modelo específico puede usar un subconjunto de las siguientes suposiciones):

1. La señal de referencia es de calidad perfecta.
2. La adaptación a la luz sigue la ley de Webber.
3. Tras la adaptación a la luz, la óptica del ojo puede ser modelada como un sistema lineal e invariante caracterizado por una PSF.
4. En el SVH existen canales de frecuencia y orientaciones selectivas en el tiempo, y la respuesta de canal puede ser modelada como un conjunto discreto de descomposiciones lineales.
5. Aunque la definición de contraste de patrones simples usados en los experimentos psicovisuales y las definiciones de contraste de imágenes naturales complejas son diferentes, son consistentes entre sí.
6. La sensibilidad al error visual entre diferentes canales de frecuencias espaciales y/o temporales pueden ser normalizadas usando un filtro paso banda o un filtro paso bajo CSF.
7. La descomposición de canal se realiza prácticamente sin pérdidas en lo referente a importancia visual, en el sentido de que la señal transformada mantiene la mayoría de la información necesaria para medir la calidad de la imagen.
8. La descomposición de canal efectivamente decorrela la estructura de la imagen, de manera que la potencia de la máscara se determina por las magnitudes (no estructuras) de los coeficientes. Después del enmascaramiento, el error percibido de cada coeficiente puede ser evaluado de forma individual.

9. Para un único coeficiente en cada canal, después de la normalización de error y el enmascaramiento, las relaciones entre la magnitud del error, $e_{l,k}$, y la distorsión percibida, $d_{l,k}$, pueden ser modeladas como una función no lineal $d_{l,k}=| e_{l,k} |^{\beta}$.
10. La distorsión general (de conjunto) percibida incrementa de forma monótona con la suma de los errores percibidos de todos los coeficientes en todos los canales.
11. Procesos de más alto nivel que tienen lugar en el cerebro humano, como reconocimiento de patrones, memoria o análisis cognitivo son menos importantes para predecir la calidad perceptual de la imagen.
12. Procesos visuales activos, como el cambio de los puntos de fijación y el ajuste adaptativo de la resolución espacial debida a la atención son menos importantes para predecir la calidad perceptual de la imagen.

Dependiendo del entorno de la aplicación, algunas de las suposiciones anteriores son válidas o razonables. Por ejemplo, en aplicaciones de compresión y comunicación de imagen y vídeo, el asumir una señal original perfecta (suposición 1) es aceptable. Sin embargo, desde un punto de vista más general, muchas de las suposiciones son discutibles y necesitan ser validadas. Existen varios problemas que son críticos a la hora de justificar la validez de los algoritmos basados en detección de error.

▪ *El problema del supraumbral*

La mayoría de los experimentos psicovisuales subjetivos se basan en el umbral de visibilidad, típicamente usando un método de opción-forzada de 2 alternativas (2AFC, *2-Alternative Forced-Choice*). Este método se usa para determinar los valores de la potencia del estímulo (también conocidos como potencia del umbral) en los cuales los estímulos comienzan a ser apreciables. Estos umbrales de medida son después usados para definir los modelos de sensibilidad del error visual, como la CSF y los diversos efectos del modelo de enmascaramiento. Sin embargo, no hay suficientes evidencias disponibles de las investigaciones de visión que apoyen la suposición de que los resultados de las medidas puedan ser generalizados para cuantificar distorsiones mucho mayores a la mínima apreciable, que es el caso para la mayoría de las aplicaciones de procesado de imagen. Esto acarrea numerosos problemas con respecto al esquema general de detección de error. Un problema es que cuando el error en un canal visual es mayor que el umbral de visibilidad, es

difícil diseñar experimentos que cumplan la suposición 9. Otro problema es concerniente a la suposición número 6, que usa el umbral de error visual mínimo apreciable para normalizar el error entre diferentes canales de frecuencia. La cuestión es: cuando los errores son mucho mayores que los umbrales, ¿pueden los errores relativos entre diferentes canales ser normalizados usando los umbrales de visibilidad?

- *El problema de la complejidad de las imágenes naturales*

La mayoría de los experimentos psicovisuales publicados son realizados usando patrones relativamente simples, como gradientes sinusoidales, áreas de Gabor, formas geométricas simples, funciones básicas de transformación o patrones de ruido aleatorio. La CSF se obtiene mediante experimentos de umbral usando patrones de frecuencia simples. Los experimentos de enmascaramiento normalmente involucran a dos (o más) patrones diferentes. Sin embargo, todos esos patrones son mucho más simples que las imágenes del mundo real, que pueden verse normalmente como una superposición de un gran número de diferentes patrones simples. ¿Son suficientes estos experimentos de patrones simples para construir un modelo que sea capaz de predecir la calidad de imágenes naturales complejas? ¿Podemos generalizar el modelo de las interacciones entre unos pocos patrones para modelar interacciones entre decenas o cientos de patrones?

- *El problema de la combinación de error Minkowski*

La tan usada fórmula de suma de error de Minkowski se basa en la diferencia entre dos señales, por lo que puede no capturar los cambios estructurales entre esas dos señales. En la figura 3.17 se proporciona un ejemplo, donde dos señales de prueba (señal 1, arriba a la izquierda, y señal 2, arriba a la derecha), se generan a partir de la señal original (arriba en el centro). La señal 1 se obtiene añadiendo un valor constante a cada punto de muestreo, mientras que los signos de la constante añadida a la señal 2 se eligen aleatoriamente entre positivo y negativo. La información estructural de la imagen original se pierde casi completamente en la señal 2, pero se mantiene muy bien en la señal 1. Para calcular la métrica de error de Minkowski, primero se resta la señal original de las señales de test, resultando las señales de error 1 y 2, que tienen estructuras muy diferentes. Sin embargo, al

aplicar el operador absoluto en las señales de error resultan exactamente la misma. Las medidas de error final de Minkowski de las dos señales de test son iguales, sin importar el valor de ß. Este ejemplo no sólo demuestra que la habilidad de "preservación de la estructura" es un factor importante al medir la similitud entre señales, sino que además muestra que la combinación de error de Minkowski es ineficiente al capturar la estructura de los errores y es una métrica con pérdidas de información estructural. Obviamente, en este ejemplo específico, el problema puede solucionarse aplicando una descomposición de canal en frecuencia espacial de las señales de error y ponderando los errores de forma diferente en cada canal mediante una CSF. Sin embargo, las señales descompuestas pueden aún mostrar diferentes estructuras en diferentes canales pudiendo tomar importancia la pérdida de información estructural de Minkowski.

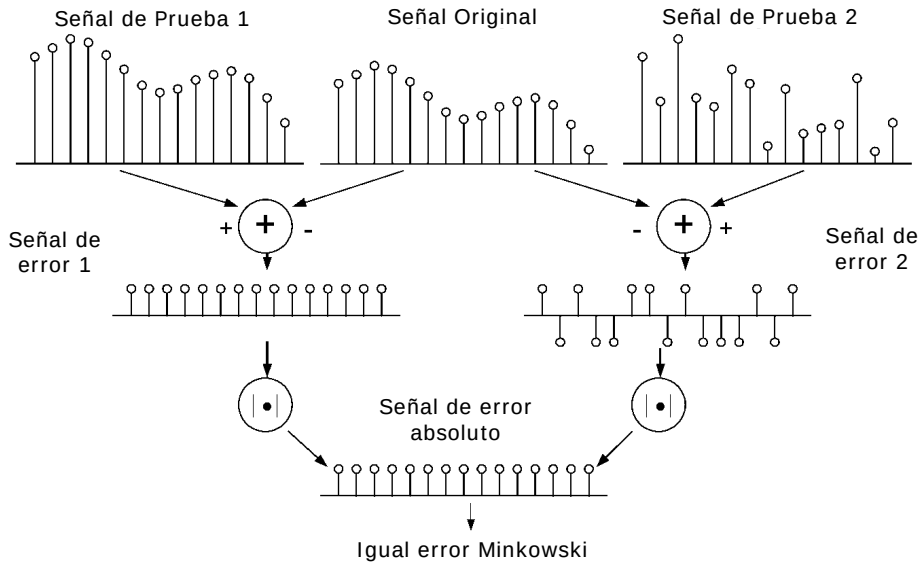


Fig. 3.17 Ilustración de la combinación de error Minkowski

- *El problema de la interacción cognitiva*

Está claro que el entendimiento cognitivo y los procesos visuales activos juegan roles importantes a la hora de evaluar la calidad de imágenes. Por ejemplo, un observador humano dará diferentes puntuaciones a una imagen dependiendo de cómo se le instruya. La información previa relativa al contenido de la imagen, o la atención, pueden afectar también la evaluación de la calidad de la imagen. Por ejemplo, en un ambiente de videoconferencia,

la diferencia entre la sensibilidad de degradación de primer plano o fondo se incrementa con la presencia del audio correspondiente a la locución de la persona del primer plano. Actualmente, la mayoría de las medidas de calidad de imagen y vídeo no consideran estos efectos. Cómo alteran estos efectos la calidad de imagen percibida, su importancia en comparación con otras características del SVH empleadas en las medidas de calidad actuales y cómo incorporar estos efectos en un modelo de medida de calidad son aspectos que todavía no se han investigado en profundidad.

3.4 Medidas basadas en distorsión estructural

El paradigma de medida de calidad de imagen y vídeo basado en detección de error considera cualquier tipo de distorsión de la escena como cierto tipo de error. Sin embargo, diferentes estructuras de error tendrán efectos distintos en la calidad percibida, por lo tanto, la efectividad de esta aproximación depende de cómo las estructuras de los errores son entendidas y representadas. La descomposición lineal de canal es la forma más comúnmente usada para descomponer las señales de error en un conjunto de componentes elementales, y los modelos de sensibilidad de error para estas componentes elementales son relativamente fáciles de obtener a partir de experimentos psicovisuales. Como se ha explicado anteriormente, debido a que los métodos de descomposición lineal de canal no pueden decorrelar completamente las estructuras de la señal, los coeficientes descompuestos muestran aún más relación con el resto. Se comprobó también como la medida de error de Minkowski no puede capturar estas correlaciones estructurales. El conocimiento actual sobre el enmascaramiento visual es aún limitado. De momento, no está claro si es posible construir un modelo de enmascaramiento comprensible, pero si lo fuera, el modelo sería muy complicado.

En este apartado se propone una forma de pensar alternativa acerca de la medida de calidad de imagen y vídeo: no es necesario considerar la diferencia entre la señal original y la distorsionada como un cierto tipo de error. Las medidas de distorsión estructural descritas a continuación resultarán en métodos de medida de calidad perceptual más eficientes y más efectivos.

3.4.1. Nueva filosofía

En 2002, Wang, Bovik y Lu [4] propusieron una nueva filosofía en el diseño de medidas de calidad de imagen y vídeo:

La función principal del sistema visual humano es extraer la información estructural del campo de visión, y el sistema visual humano está altamente adaptado para este propósito. Por lo tanto, una medida de la distorsión estructural podría ser una buena aproximación de la distorsión de imagen percibida.

La nueva filosofía puede entenderse mejor en comparación con la filosofía basada en detección de error:

En primer lugar, la diferencia más importante de la nueva filosofía con respeto a la basada en detección de error es el cambio de medida de error a medida de distorsión estructural. Aunque el error y la distorsión estructural en ocasiones concuerdan, en muchas circunstancias la misma cantidad de error puede resultar en distorsiones estructurales significativamente diferentes. Un buen ejemplo es el que se proporciona en las figuras 3.18 y 3.19, donde la imagen original "Lena" se altera con una amplia variedad de distorsiones: ruido impulsivo sal y pimienta, ruido aditivo gaussiano, ruido multiplicativo, cambio de media, expansión de contraste, desenfoque y codificación JPEG muy agresiva. Se han ajustado todas las imágenes para obtener el mismo MSE en ellas (relativo a la original), excepto por la comprimida en JPEG, que tiene un MSE menor. Es interesante ver que las imágenes con casi idéntico MSE tienen calidad perceptual drásticamente diferente. La evaluación subjetiva muestra que la expansión de contraste y el cambio de media proporcionan imágenes de gran calidad perceptual, mientras que el desenfoque y la compresión JPEG tienen las puntuaciones más bajas. Esto no sorprende, al comprender la nueva filosofía, ya que los cambios estructurales de la original a las imágenes con expansión de contraste o cambio de media son triviales, pero en la imagen desenfocada y en la comprimida la modificación estructural es muy significativa.

Fig. 3.18 Evaluación de las imágenes “Lena” con distintos tipos de ruido:
- Arriba a la izquierda: imagen original “Lena”, 512 x 512, 8 bits/píxel
- Arriba a la derecha: imagen con ruido impulsivo sal y pimienta, MSE=225, Q=0,6494
- Abajo a la izquierda: imagen con ruido aditivo gaussiano, MSE=225, Q=0,3891
- Abajo a la derecha: imagen con ruido multiplicativo, MSE=225, Q=0,4408

En segundo lugar, otra diferencia importante de la nueva filosofía es que considera la degradación de la imagen como pérdida de información estructural percibida. Por ejemplo, en la figura 3.19, la expansión de contraste tiene una mejor calidad que la compresión JPEG simplemente porque casi toda la información estructural de la imagen original se preserva, en el sentido de que la imagen original puede ser recuperada mediante una simple transformación lineal inversa punto a punto de luminancia. Aparentemente, gran cantidad de información de la imagen original se ha perdido completamente en la imagen

comprimida. La razón de que una medida de pérdida de información estructural pueda ser considerada como predicción de la percepción visual se basa en la suposición de que el SVH funciona de forma similar –está adaptado a extraer información estructural y detectar cambios en ella. Por el contrario, una aproximación basada en detección de error estima los errores percibidos para representar la degradación de la imagen. Si esto fuera correcto, entonces se debería observar una distorsión perceptual importante para la imagen con expansión de contraste porque existe una gran diferencia (en términos de error) con la imagen original.

Fig. 3.19 Evaluación de las imágenes “Lena” con distintos tipos de distorsiones:
- Arriba a la izquierda: imagen con cambio de media, MSE=225, Q=0,9894
- Arriba a la derecha: imagen con extensión de contraste, MSE=225, Q=0,9372
- Abajo a la izquierda: imagen desenfocada, MSE=225, Q=0,3461
- Abajo a la derecha: imagen comprimida JPEG, MSE=215, Q=0,2876

En tercer lugar, la nueva filosofía emplea una aproximación *top-down*, que comienza desde un nivel muy alto –simulando la funcionalidad hipotética de conjunto del SVH. En comparación, la filosofía basada en detección de error, usa una aproximación *bottom-up*, que intenta simular la función de cada componente relevante del sistema visual humano y combinarlas juntas, con la intención de que la combinación se comporte de forma similar a como lo hace el conjunto del SVH.

Cómo aplicar la nueva filosofía para crear medidas concretas de calidad de imagen y vídeo es un trabajo abierto y en desarrollo. Hay diferentes implementaciones, dependiendo de cómo se interpretan y cuantifican los conceptos de información estructural y distorsión estructural. De forma general, hay dos formas de implementar un algoritmo de medida de calidad usando esta nueva filosofía. La primera es desarrollar un entorno de descripción de características de las imágenes naturales, que cubre la mayoría de la información estructural de una imagen. En ese entorno, los cambios en la información estructural entre la señal original y la distorsionada pueden ser cuantificados. La segunda es diseñar un método de comparación de estructuras que pueda comparar la similitud o la diferencia estructural entre las imágenes de forma directa.

3.4.2. Aproximación para el indexado de calidad de imágenes

Como un primer intento de implementar esta nueva filosofía, a continuación se presenta una sencilla aproximación de indexado de calidad de imagen realizado por Wang y Bovik [50]. A pesar de no evaluar secuencias de vídeo, la explicación de este algoritmo sentará las bases de las métricas estructurales de vídeo descritas en el siguiente capítulo.

Sean $x = \{x_i \mid i = 1,2,\cdots N\}$ e $y = \{y_i \mid i = 1,2,\cdots N\}$ las señales de las imágenes original y de prueba, respectivamente. El índice de calidad propuesto es:

$$Q = \frac{4\sigma_{xy}\,\bar{x}\,\bar{y}}{\left(\sigma_x^2 + \sigma_y^2\right)\left[\left(\bar{x}\right)^2 + \left(\bar{y}\right)^2\right]},$$

donde

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \ , \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \ ,$$

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \ , \quad \sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

El rango dinámico de Q es [–1, 1]. El mejor valor 1 se alcanza si y sólo si $y_i=x_i$ para todo $i=1,2,\cdots N$. El valor más bajo –1 ocurre cuando $y_i = 2\bar{x} - x_i$, para todo $i=1,2,\cdots N$.

Este índice de calidad modela cualquier distorsión como combinación de tres factores: pérdida de correlación, distorsión de media y distorsión de contraste. Para entender esto, se reescribe la definición de Q como el producto de tres componentes:

$$Q = \frac{\sigma_{xy}}{\sigma_x \, \sigma_y} \cdot \frac{2\,\bar{x}\,\bar{y}}{(\bar{x})^2 + (\bar{y})^2} \cdot \frac{2\,\sigma_x \, \sigma_y}{\sigma_x^2 + \sigma_y^2},$$

La primera componente es el coeficiente de correlación entre ***x*** e ***y***, que mide el grado de correlación lineal entre las imágenes, y su rango dinámico es [–1, 1]. El mejor valor 1 se obtiene cuando $y_i = a\,x_i + b$ para todo $i=1,2,\cdots N$, donde a y b son constantes y $a>0$. Se considera la correlación lineal como un factor muy importante en la comparación de estructuras de dos señales. De hecho, una señal modificada linealmente punto a punto puede ser recuperada exactamente con una simple transformación lineal inversa punto a punto. En este sentido, la información estructural se preserva. Además, un decrecimiento en el coeficiente de correlación lineal proporciona una medida cuantitativa de cómo la señal cambia de forma no lineal. Obviamente, incluso si ***x*** e ***y***, están relacionadas linealmente, puede haber distorsiones relativas entre ellas, lo que se evalúa en la segunda y tercera componente. La segunda componente con un rango de [0, 1] mide cómo de parecidos son los valores medios de ***x*** e ***y***. Es igual a 1 si y sólo si $\bar{x} = \bar{y}$. σ_x y σ_y pueden ser vistas como estimaciones burdas del contraste de ***x*** e ***y***, por lo tanto la tercera componente mide cómo de similares son los contrastes de las imágenes. Su rango de valores es también [0, 1], donde el mejor valor 1 se alcanza si y sólo si $\sigma_x=\sigma_y$.

Las señales de imágenes son generalmente no estacionarias y la calidad de la imagen es a menudo variante espacialmente. En la práctica, normalmente se desea evaluar una imagen completa usando un único valor de calidad de conjunto. Por ello, es razonable medir características estadísticas localmente y luego combinarlas juntas. Este método de medida se aplica a regiones locales usando una aproximación de ventana deslizante. Empezando desde la esquina superior izquierda, una ventana de tamaño BxB se mueve píxel por píxel horizontal y verticalmente a través de todas las filas y todas las columnas de la imagen hasta que se alcanza la esquina inferior derecha. En el paso *j-ésimo,* el índice de calidad local Q_j se calcula para la ventana en cuestión. Si hay un total de M pasos, entonces el índice de calidad total viene dado por:

$$Q = \frac{1}{M}\sum_{j=1}^{M} Q_j$$

Se ha demostrado que muchos algoritmos de medida de calidad de imagen funcionan de forma consistente si las imágenes distorsionadas que se comparan han sido creadas a partir de la misma imagen original y el mismo tipo de distorsiones. De hecho, para esas comparaciones, la MSE o PSNR es normalmente suficiente para producir evaluaciones de calidad útiles. Sin embargo, la efectividad de los modelos de evaluación de la calidad se degrada significativamente cuando los modelos son empleados para comparar la calidad de imágenes distorsionadas de distintos originales o con diversos tipos de distorsión. Por ello, pruebas de imágenes cruzadas y distorsiones cruzadas son muy útiles al evaluar la efectividad de una métrica de calidad.

Las imágenes de las figuras 3.18 y 3.19 son buenos ejemplos para probar la capacidad de distorsión cruzada de los algoritmos de medida de calidad. Obviamente, el MSE proporciona unos resultados muy pobres en ese caso. Sin embargo, podemos observar los índices de calidad calculados para cada una de de las imágenes, fijando la ventana deslizante a B=8. Los resultados muestran una sorprendente consistencia con las evaluaciones subjetivas.

Fig. 3.20 Evaluación de calidad de imagen con desenfoque:
- Arriba a la izquierda: imagen original “woman”
- Arriba a la derecha: imagen “woman” desenfocada, MSE=200, Q=0,3483
- Medio a la izquierda: : imagen original “man”
- Medio a la derecha: imagen “man” desenfocada MSE=200, Q=0,4123
- Abajo a la izquierda: imagen original “barbara”
- Abajo a la derecha: imagen “barbara” desenfocada, MSE=200, Q=0,6594

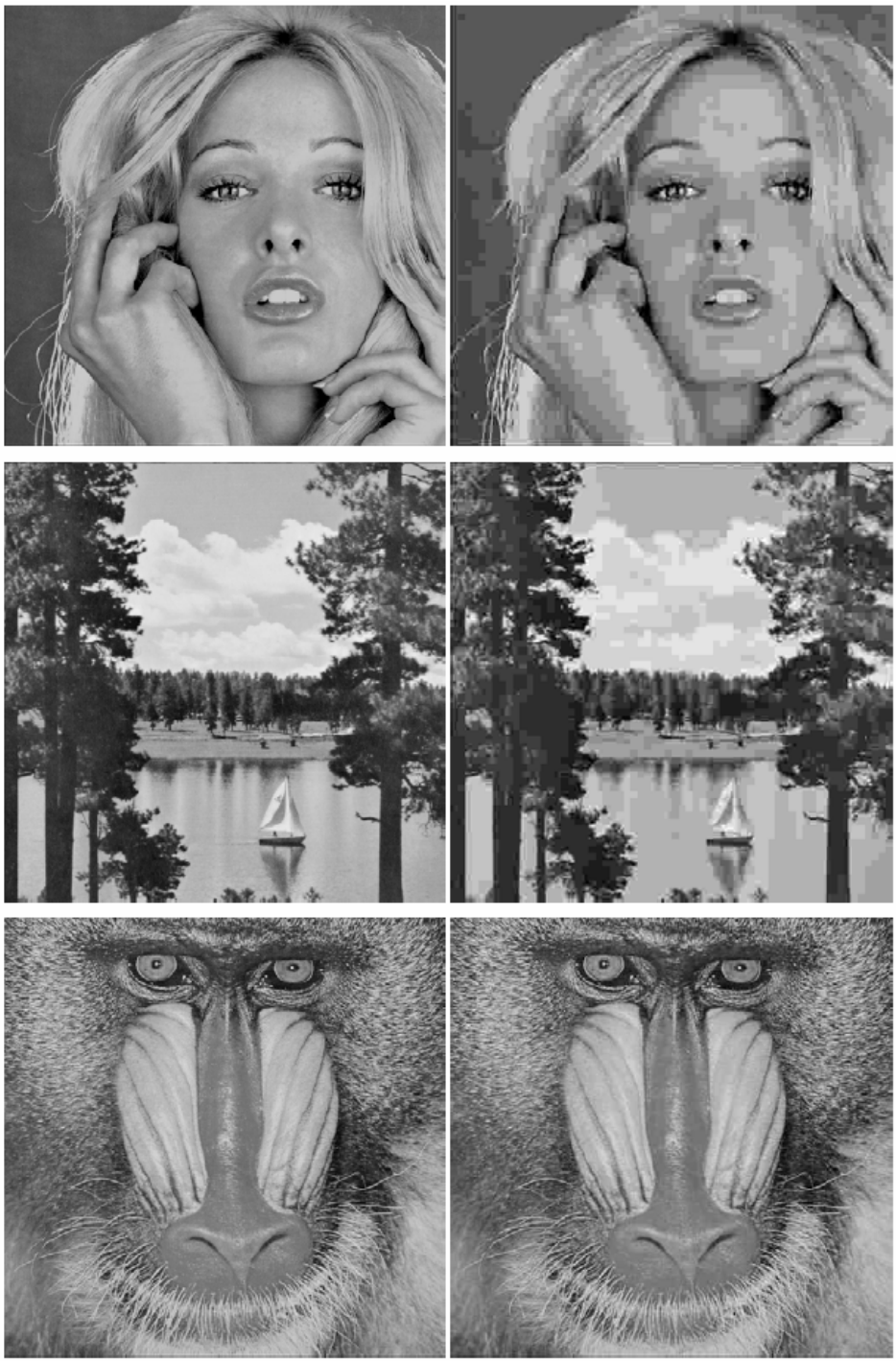


Fig. 3.21 Evaluación de calidad de imagen con compresión JPEG:
- Arriba a la izquierda: imagen original “tiffany”
- Arriba a la derecha: imagen “tiffany” comprimida, MSE=165, Q=0,3709
- Medio a la izquierda: : imagen original “lake”
- Medio a la derecha: imagen “lake” comprimida MSE=167, Q=0,4606
- Abajo a la izquierda: imagen original “mandrill”
- Abajo a la derecha: imagen “mandrill” comprimida, MSE=163, Q=0,7959

En las figuras 3.20 y 3.21, se emplean diferentes imágenes con el mismo tipo de distorsión para probar la capacidad de imágenes cruzadas del índice de calidad. En la figura 3.20, se desenfocan tres imágenes diferentes, de forma que tienen el mismo MSE con respecto a sus originales. En la figura 3.21, otras tres imágenes son comprimidas usando JPEG, y los pasos de cuantificación del JPEG se seleccionan de forma que las tres imágenes comprimidas tengan un MSE similar en comparación con sus originales. De nuevo, el MSE tiene una correlación muy pobre con la calidad percibida en las pruebas, y el algoritmo de indexado propuesto proporciona una consistencia mayor con las evaluaciones visuales.

El método de indexación por calidad propuesto es sólo una implementación rudimentaria del nuevo paradigma. Aunque proporciona resultados prometedores en las actuales condiciones de prueba limitadas, se necesitan experimentos más extensos para validar y optimizar el algoritmo. Se necesitan establecer más conexiones teóricas y experimentales con respecto a la percepción visual humana.

Como se ha comentado, sirve como base para algunos algoritmos de medida de calidad de vídeo. En [51], Wang, Lu y Bovik trataron de calcular el índice fotograma a fotograma para una secuencia de vídeo y combinarlo con otras características de distorsión como el efecto de bloques para generar una medida de calidad perceptual de vídeo.

4

Algoritmos analizados

Tras haber estudiado en el apartado anterior los distintos tipos de algoritmos para el cálculo objetivo de la calidad subjetiva de secuencias de vídeo, a continuación se analizan tres métodos ampliamente conocidos y usados. Se expondrán sus principales características y limitaciones con el fin de servir de base al algoritmo de evaluación propuesto. Su elección se debe a las características específicas de cada uno de ellos: su importancia (el segundo de los métodos analizados se ha incluido como medida de referencia en dos recomendaciones de la Unión Internacional de Telecomunicaciones [52, 53]), los resultados obtenidos en evaluaciones previas, así como a la disponibilidad de implementaciones por parte de los autores con las que evaluar su rendimiento.

El primer método es la Medida de Calidad de Vídeo (VQM, por sus siglas en inglés) desarrollada por la Administración Nacional de Telecomunicación e Información (NTIA), dependiente de la Cámara de Comercio de los Estados Unidos. De los 5 modelos de que dispone la medida, se estudiará el modelo "general", destinado a un amplio rango de distorsiones y tipos de secuencias de referencia. Esta medida es, actualmente, la que proporciona mejores resultados. Su principal inconveniente es la carga computacional, que hace extremadamente compleja su implantación en sistemas de tiempo real.

Posteriormente se detallará la Medida de Calidad de Vídeo Digital de Watson modificada (DVQ, por sus siglas en inglés). Se basa en el modelo DVQ de Watson, que usa la Transformada Discreta del Coseno. Aunque este algoritmo no ofrece unos resultados muy buenos, se elige por su sencillez y como muestra de la incorporación de aspectos de la sensibilidad visual humana en un sencillo algoritmo basado en detección de error.

Por último se analizará el algoritmo VSSIM o Índice de Similitud Estructural de Vídeo, que hace uso del Índice SSIM de imágenes estáticas. Esta medida, basada en distorsión estructural, será base fundamental del algoritmo propuesto.

Aunque existen numerosas propuestas en lo que a métricas de evaluación de calidad de vídeo se refiere, se han elegido estas tres por las razones anteriormente expuestas. Podemos encontrar algoritmos que ofrecen resultados muy pobres, en otros casos existen propuestas no concretadas en ninguna implementación disponible, y también existen estudios más actuales que aún se encuentran en fase de desarrollo.

4.1 Medida de Calidad de Vídeo NTIA

La medida de calidad (VQM) de la NTIA [56] tiene 5 modelos que extraen y comparan diferentes parámetros del vídeo. A continuación se describirá el modelo "general" que está diseñado para ser una medida de propósito general en sistemas de vídeo que abarquen un amplio rango de calidades y tasas de bit. Este modelo usa tecnología de referencia reducida (RR) y proporciona estimaciones de la calidad global del vídeo.

El modelo general y sus técnicas de calibración asociadas abarcan un sistema de medida de calidad de vídeo completamente automático. En la figura 4.1 de la página siguiente se muestra un diagrama general del sistema.

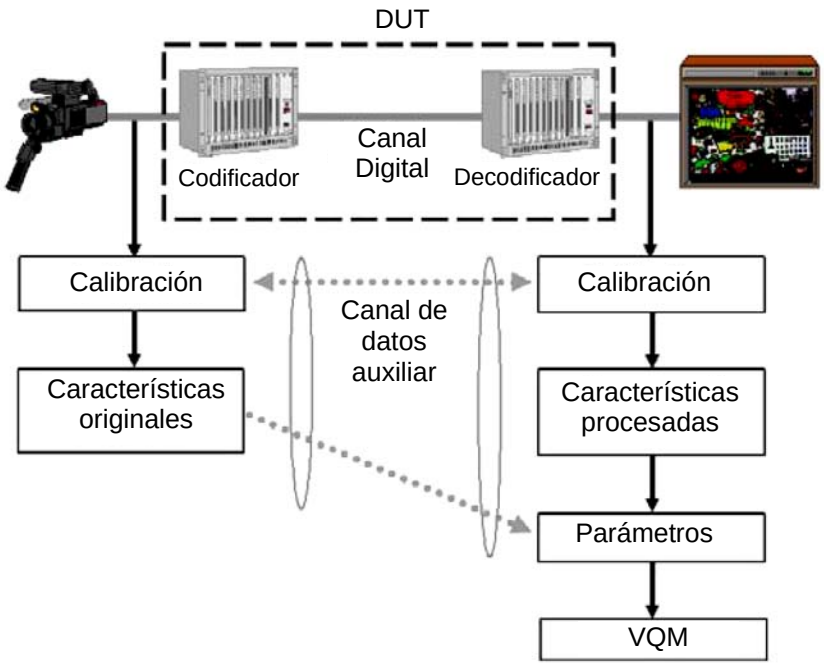


Fig. 4.1 Diagrama esquemático del algoritmo NTIA

La calibración de la señal original y la procesada incluye alineación espacial, estimación de región válida, cálculo de la compensación de ganancia y nivel, y alineación temporal. El cálculo del VQM implica la extracción de características perceptuales, procesado de parámetros de calidad de vídeo, y combinación de parámetros para construir el modelo general.

4.1.1 Alineación espacial

El proceso de alineación espacial determina la desviación horizontal o vertical del vídeo procesado. La precisión del algoritmo de alineación espacial es de 0,5 píxeles para desplazamientos horizontales y de una línea en desviaciones verticales. Tras calcular la alineación espacial, el desplazamiento se elimina de la señal de vídeo procesada.

La alineación espacial debe ser calculada antes de la región válida de procesado (PVR, por sus siglas en inglés). La PVR se define como el área de la imagen de vídeo procesada que contiene información válida. Por otro lado, la alineación espacial ha de tener lugar antes de la alineación temporal.

Todas estas correcciones han de realizarse comparando la secuencia procesada con la original. Si el vídeo procesado estuviera espacialmente desplazado con respecto a la referencia y no se corrigiera esa desviación, el resto de calibraciones y medidas estarían corruptas. Sin embargo, la alineación espacial no puede determinarse correctamente a menos que la PVR, la compensación de ganancia y nivel y la alineación temporal sean también conocidas. La interdependencia de estas medidas produce un círculo vicioso, lo que genera un problema a la hora de estimar los valores de calibración. El cálculo de la alineación espacial para un campo procesado requiere conocer la PVR, la compensación de ganancia y nivel, así como la alineación temporal. Sin embargo, estas cantidades no se pueden determinar hasta que la alineación espacial se realiza. Una búsqueda exhaustiva sobre todas las variables requeriría un número inmenso de cálculos. Localizar la búsqueda en una zona muy restringida liberaría coste computacional pero podría dar lugar a errores de alineación. La solución presentada aquí realiza una búsqueda iterativa para encontrar el mejor ajuste posible entre fotogramas originales y procesados.

En primer lugar se realiza una estimación inicial de referencia para desplazamientos horizontales, verticales y temporales en un fotograma, siguiendo una búsqueda en varios pasos. El primer paso consiste en una búsqueda amplia en un conjunto determinado de desplazamientos espaciales, cuyo objetivo es acercarse a un ajuste correcto con respecto al fotograma original. Aún no se considera la compensación de ganancia y el PVR se fija de forma que se excluya la zona de la imagen que pueda sufrir sobreescaneo (área de la imagen que no será visible en determinados dispositivos de visualización). Con dicho PVR se evitará usar zonas inválidas del vídeo. El segundo paso es otra búsqueda amplia para aproximar el desplazamiento espacial que se realiza usando un rango limitado de fotogramas originales. El tercer paso realiza una búsqueda espacio-temporal localizada para ajustar de forma precisa la estimación temporal y espacial. Cada búsqueda fina incluye un conjunto pequeño de desplazamientos espaciales centrados alrededor de la estimación actual y tres fotogramas para la alineación temporal. La condición de desplazamiento cero se incluye como comprobación de seguridad que ayuda a prevenir que el algoritmo converja en un mínimo local. Este tercer paso se itera hasta 5 veces. Si las búsquedas finas no consiguen encontrar un resultado estable, la búsqueda anterior se repite utilizando un fotograma procesado distinto al empleado. Este proceso produce una estimación de referencia que será actualizada de forma periódica.

El algoritmo de alineación espacial realiza la estimación en un fotograma procesado a una determinada frecuencia (por ejemplo, un fotograma cada medio segundo). Usando la estimación de referencia como punto de partida, el algoritmo realiza búsquedas finas (como se ha explicado anteriormente) y estimaciones de la compensación de ganancia y nivel.

Para algunos fotogramas, el algoritmo de alineación espacial puede fallar. Normalmente, cuando el desplazamiento espacial no se estima correctamente, se debe a características de la escena. Consideremos una escena creada digitalmente con un movimiento de cámara progresivo a la izquierda. Como la escena se ha creado digitalmente, puede tener un movimiento horizontal de exactamente dos píxeles por fotograma. Desde el punto de vista del algoritmo, sería imposible diferenciar entre la alineación espacial correcta usando el fotograma original, o un desplazamiento horizontal de dos píxeles usando el fotograma original previo. Otro ejemplo: en una imagen consistente en negro puro con líneas verticales blancas, el desplazamiento vertical es totalmente ambiguo.

El algoritmo descrito requiere un canal de datos auxiliar relativamente alto, debido a la comparación píxel a píxel entre los fotogramas originales y procesados. Esto podría afectar al diseño de una aplicación viable de monitorado de calidad. Afortunadamente, cada equipo de transmisión de vídeo (codificador, decodificador, transmisión analógica, etc.) introducirá normalmente un desplazamiento espacial constante; con lo que la información necesaria para la estimación del desplazamiento será requerida únicamente en la configuración inicial del sistema.

4.1.2. Región válida de procesado

Las secuencias de vídeo pueden poseer un borde de píxeles y líneas que no contienen imagen válida. Un sistema de vídeo digital que utilice compresión puede además reducir el área de la imagen con el fin de almacenar bits de transmisión. Adicionalmente, algunos sistemas de procesado pueden deteriorar algunas líneas o píxeles. Si éstos tienen lugar en el área de sobreescaneo, el usuario final normalmente no apreciará los errores; sin embargo si son suficientemente grandes pudiera ocurrir que el usuario observara un borde negro alrededor de las imágenes. Para evitar que estas áreas influyan en la VQM, se excluyen de

la medida. Debido a que el comportamiento de algunos sistemas es dependiente de la señal de vídeo, la PVR se calcula para cada escena. Una vez que se tiene la región válida, los píxeles no válidos se descartan tanto de la secuencia original como de la procesada.

El algoritmo usado comienza suponiendo que los bordes externos de cada fotograma procesado contienen vídeo no válido. El tamaño de esta región se fija empíricamente, basándose en observaciones de los sistemas de vídeo actuales. Para sistemas de 525 líneas, la región no válida abarca 6 píxeles a izquierda y derecha, 6 líneas en la parte superior, y 4 líneas en la parte inferior. El algoritmo PVR comienza fijando la zona no válida a esta configuración por defecto. Tras ello se examinan los píxeles inmediatamente posteriores en la actual zona válida. Si el valor medio del píxel es negro o sube progresivamente desde negro (algunos bordes efectúan un degradado desde negro), entonces se disminuye el tamaño de la región válida de procesado. Repitiendo este cálculo, la PVR disminuye iterativamente.

Las condiciones de parada pueden verse afectadas por el contenido de la escena. Por ejemplo, una imagen que contiene negro puro a la izquierda llevará al algoritmo a pensar que la región válida de procesado está más a la derecha de lo que realmente está. Por esta razón, el algoritmo se aplica a varias imágenes de la secuencia procesada y para la estimación final se usa la PVR mayor (con un margen de seguridad añadido).

Hay que tener en cuenta que el algoritmo se ha realizado con un diseño conservador: un examen manual de la escena produciría una PVR más amplia en la mayoría de los casos. Esto se debe a que el descarte de una pequeña cantidad de vídeo tendrá poco impacto en la estimación de la calidad; sin embargo, la inclusión de vídeo corrupto en los cálculos de calidad puede tener un impacto mucho mayor en la estimación.

4.1.3. Compensación de ganancia y nivel

Un requisito previo para realizar estos cálculos es que las imágenes originales y procesadas han de estar alineadas tanto espacial como temporalmente. El método usado asume que las señales Y, Cb y Cr tienen un cambio de ganancia y nivel independiente. Sin

embargo, en sistemas de video compuesto o S-vídeo, es posible tener una rotación de fase en la información de crominancia ya que estas dos componentes se multiplexan en una señal compleja como un vector con módulo y fase. El algoritmo que se presenta no calibrará adecuadamente los sistemas que introduzcan una rotación de fase en la información de crominancia (por ejemplo el ajuste de matiz de un televisor).

Las regiones válidas de los fotogramas original y procesado se dividen en pequeños bloques cuadrados o sub-regiones. En cada uno de ellos se calcula la media en el espacio de las muestras [Y Cb Cr] para formar imágenes espacialmente sub-muestreadas. Para registrar temporalmente un fotograma procesado (con desplazamiento espacial fijado), se calcula la desviación típica de cada imagen diferencia (original menos procesada) usando las muestras Y de los fotogramas sub-muestreados. Para un fotograma procesado concreto, el desplazamiento temporal que produce la menor desviación típica (la mayor cancelación con respecto al original) se elige como la mejor opción. Se realiza entonces un ajuste lineal para procesar la desviación de ganancia entre los fotogramas sub-muestreados (originales y procesados). Este ajuste lineal se aplica de forma independiente a cada uno de los tres canales: Y, Cb y Cr.

Aunque la compensación de ganancia y nivel se calcula para los canales de crominancia, estos factores de corrección no se aplican. El modelo general utiliza únicamente la compensación de ganancia y nivel de la luminancia (o canal Y). Los cambios en las componentes Cb y Cr se consideran distorsiones por los cuales el dispositivo bajo estudio debería ser penalizado.

4.1.4. Alineación temporal

Los actuales sistemas de vídeo digital requieren por lo general décimas de segundo para procesar y transmitir el vídeo desde el envío de la cámara hasta la recepción en el dispositivo de visualización. Un retardo de vídeo excesivo impide una comunicación bidireccional efectiva. Por lo tanto, son importantes métodos objetivos para medir el retardo extremo a extremo de las comunicaciones tanto para usuarios como para proveedores de servicio (especificaciones y comparación de servicios interactivos). El retardo depende tanto de atributos dinámicos de la escena original (como detalle espacial o movimiento)

como de los sistemas de vídeo (tasa de bits). Por ejemplo, escenas con gran cantidad de movimiento pueden sufrir más retardo que las escenas con menor cantidad. Por ello, las medidas de retardo se realizan a tiempo real para que sean representativas y precisas. Se necesitan estimaciones del retardo de vídeo para alinear temporalmente las señales de vídeo original y procesado antes de realizar medidas de calidad.

Algunos sistemas de transmisión de vídeo pueden proporcionar información de sincronización temporal (por ejemplo, fotogramas originales y procesados pueden etiquetarse con algún tipo de información temporal). Sin embargo, por lo general la sincronización entre las señales de vídeo original y procesada ha de ser medida. Aquí se presenta una técnica para estimar el retardo del vídeo a nivel de fotograma. Funciona correlando imágenes de resolución menor, submuestreadas en el espacio y extraídas de los vídeos original y procesado. Esta técnica estima el retardo individual de cada fotograma y posteriormente los combina para calcular el retardo medio de la secuencia completa.

Para reducir la influencia de las distorsiones en la alineación temporal, las imágenes originales y procesadas se normalizan a varianza unidad después de ser submuestreadas. Cada fotograma procesado se compensa temporalmente usando la técnica descrita en el algoritmo de ganancia y nivel (encontrar el fotograma original que minimice la desviación típica de a diferencia entre original y procesado). Esto localiza la imagen original más parecida para cada imagen procesada.

Sin embargo, lo importante no es identificar el fotograma original del que proviene cada fotograma procesado, sino calcular el retardo relativo entre las secuencias (en segundos o fotogramas). Las medidas de retardo de una serie de imágenes se combinan en un histograma, que se suaviza posteriormente. Si en algún extremo del histograma, hay una acumulación de medidas, entonces el intervalo de incertidumbre usado fue demasiado pequeño y el algoritmo se ejecuta de nuevo con un valor mayor. La mejor alineación temporal de la escena coincide con el máximo del histograma suavizado.

Al contrario que los algoritmos de calibración previos, en la alineación temporal se examina cada fotograma procesado. Algunas de las estimaciones serán incorrectas, pero estos errores tenderán a distribuirse de forma aleatoria, con lo que el resultado final no se verá muy afectado.

4.1.5. Descripción general de las características y cálculo de parámetros

Una característica de calidad en el contexto de este algoritmo se define como una cantidad de información asociada con, o extraída de, una subregión espacial o temporal de una señal de vídeo (original o procesada). Los flujos de características son una función en el espacio y en el tiempo que caracterizan los cambios perceptuales en las propiedades espaciales, temporales y de crominancia de la señal de vídeo. Comparando las características extraídas del vídeo procesado calibrado con las extraídas del vídeo original, se pueden calcular parámetros de calidad que son indicativos de los cambios perceptuales en la calidad de la secuencia.

De forma conceptual, todas las características usadas en el modelo general, llevan a cabo los mismos pasos. Se aplica un filtro perceptual al vídeo para mejorar e intensificar algunas propiedades de la calidad percibida, como información de bordes. Tras ello, se extraen características de las subregiones espacio-temporales usando una función matemática (por ejemplo, la desviación típica). Por último, se aplica un umbral de percepción a las características extraídas. Todas las características operan de forma independiente al tamaño del vídeo, por ejemplo, la región espacio-temporal no cambia al modificar el tamaño de la imagen.

Cada filtro perceptual distingue algún aspecto de calidad del vídeo. El plano de luminancia contiene información sobre bordes y ruido. Un realce de bordes en el plano de luminancia identificará de forma más precisa desenfoques, efectos de bloque, y otros efectos a gran escala. Los planos de color Cb y Cr son útiles para identificar distorsiones de matiz y errores digitales de transmisión.

Después de que las señales de vídeo original y procesada hayan sido filtradas perceptualmente, éstas se dividen en regiones espacio-temporales adyacentes. Las regiones quedan descritas mediante (1) el número de píxeles horizontalmente, (2) el número de líneas verticalmente, y (3) la duración temporal de la región. Al haber calibrado el vídeo procesado, para cada una de sus regiones, existe su correspondiente región espacio-temporal original. De ellas se extraen las características usando una función matemática simple. Las

dos funciones que proporcionan mejores resultados son la media (que ofrece el valor de píxel medio) y la desviación típica, que estima la dispersión de los valores.

Por último, algunas características se recortan para evitar medir distorsiones que no son percetibles. El recorte es de la forma:

$$f_{clip} = \max(f, umbral)$$

donde f es la característica antes del recorte, el umbral es el valor de recorte a partir del cual la característica ya no es perceptible y f_{clip} es la característica tras el recorte. Gracias a esta operación se reduce la sensibilidad a distorsiones no perceptibles.

Mientras que las características de calidad cuantifican algún aspecto perceptual de una señal de vídeo, los parámetros de calidad comparan las características originales y procesadas para obtener una medida general de la distorsión del vídeo. Conceptualmente, todos los parámetros usados por el modelo general siguen los mismos pasos (como ocurría con las características). Primero, el valor de cada característica de cada región procesada se compara con su correspondiente valor original usando funciones de comparación que simulan la percepción de distorsión. Posteriormente, se aplican funciones de combinación preceptúales a lo largo del espacio y del tiempo. Con ello se conseguirán valores únicos en los parámetros de calidad para la secuencia completa, que normalmente tendrá una duración de 8 a 10 segundos. Los valores finales pueden también ser escalados y recortados para tener en cuenta las no-linealidades y ajustar mejor los parámetros a la sensibilidad humana de las distorsiones en vídeo.

La distorsión perceptual de cada región se calcula usando funciones de comparación que han sido desarrolladas para modelar el enmascaramiento visual de las distorsiones espaciales y temporales. Algunas características usan una función de comparación que utiliza una distancia euclídea entre las características originales y procesadas:

$$p = \sqrt{\left(f_o - f_p\right)^2 + \left(f_{o2} - f_{p2}\right)^2}$$

Sin embargo, la mayoría usa una comparación racional o logarítmica:

$$p = \frac{(f_p - f_o)}{f_o} \qquad p = \log 10\left(\frac{f_p}{f_o}\right)$$

donde f_o y f_{o2} son valores de características originales, y f_p y f_{p2} son los valores de las correspondientes características procesadas.

Estas funciones implican que la percepción de la distorsión es inversamente proporcional a la cantidad de actividad temporal o espacial que está presente. En otras palabras, las distorsiones espaciales se vuelven menos visibles a medida que la actividad espacial incrementa (enmascaramiento espacial) y las distorsiones espaciales se vuelven menos visibles a medida que la actividad temporal incrementa.

La comparación logarítmica y la racional producen una mezcla de valores positivos y negativos, donde los valores positivos indican ganancia y los valores negativos indican pérdidas.

Los parámetros de las regiones espacio-temporales forman vectores tridimensionales a lo largo del eje temporal y las dos dimensiones espaciales. Para el paso de combinación, se asocian en un único valor las distorsiones de las regiones con el mismo índice temporal usando alguna función como la media, la desviación típica o percentil; dando lugar a una sucesión temporal de valores.

La sucesión temporal de parámetros se combina de nuevo usando una función de asociación temporal para producir un parámetro objetivo para la secuencia completa de vídeo.

Por último, se puede aplicar una función de recorte para reducir la sensibilidad del parámetro a pequeños valores de distorsión. Esta función de recorte para parámetros positivos es:

$$p' = \begin{cases} 0 & si \quad p \le t \\ p - t & en \ otro \ caso \end{cases}$$

donde t es el umbral.

4.1.6. Parámetros del modelo general

El modelo general contiene siete parámetros independientes. Cuatro de ellos se basan en características espaciales extraídas del plano de luminancia Y; otros dos se basan en características del vector formado por las dos componentes de crominancia (Cb, Cr), y el último se basa en el producto de características que evalúan el contraste y el movimiento, ambas extraídas de la componente de luminancia. Los siete parámetros se procesan como se describe a continuación:

A. *Parámetro "si_loss"*

Este parámetro detecta una pérdida o disminución de información espacial (debida, por ejemplo, a un desenfoque). Usa un filtro espacial de 13 píxeles (SI13) que fue desarrollado para medir distorsiones de bordes perceptualmente significantivos. Un método alternativo para extraer bordes es el filtro de Sobel, pero un filtro de 3 píxeles de Sobel detectaría detalles tan finos que la gente no notaría. El SI13 usa dos máscaras de filtrado de 13x13 píxeles que se crean replicando horizontal y verticalmente el siguiente vector:

[-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]

Los filtros vertical y horizontal se aplican de forma separada al plano de luminancia. Las imágenes filtradas resultantes (IH e IV) se combinan en una única imagen (ISI13) usando la distancia euclídea (raíz cuadrada de la suma de los cuadrados).

El parámetro *si_loss* se calcula llevando a cabo los siguientes pasos:

1. Aplicar el filtro SI13 al plano de luminancia.
2. Dividir cada secuencia de vídeo en regiones espacio-temporales de 8 píxeles x 8 líneas x 0,2 segundos (región óptima hallada experimentalmente).
3. Procesar la desviación típica de cada región.
4. Aplicar un umbral de percepción, reemplazando valores menores de 12 por 12.
5. Comparar los flujos de características originales y procesadas usando la función de comparación racional.
6. Reducción espacial calculando la media del 5% de muestras peores (más distorsionadas) de las regiones espacio temporales de duración 0,2 segundos.

7. Reducción temporal ordenando los valores en el tiempo y seleccionando el percentil 10. Al ser los valores negativos, de esta forma se procesa temporalmente los peores casos.

B. *Parámetro "hv_loss"*

Este parámetro detecta desviaciones de los bordes horizontal o vertical hacia la diagonal, como podría ocurrir por ejemplo si los bordes horizontales y verticales sufren más desenfoque que los bordes diagonales. Usa las imágenes filtradas horizontal y verticalmente que producía el filtro SI13. Sin embargo, ahora se crean dos nuevos filtros perceptuales: uno que contiene los bordes horizontales y verticales (HV) y otro que contiene los bordes diagonales (HVBAR). Se procesa un ángulo de bordes para cada píxel tomando la arcotangente del cuarto cuadrante de los valores H y V filtrados. La imagen HV contiene valores donde el ángulo está entre 0,225 radianes (horizontal y vertical) y cero. La imagen HVBAR contiene valores donde el ángulo indica un borde diagonal. Los píxeles con un valor menor que 20 no se usan (se reemplazan por cero), porque el cálculo del ángulo es poco fiable.

El parámetro *hv_loss* se calcula mediante los siguientes nueve pasos:

1. Aplicar los filtros perceptuales HV y HVBAR a cada plano de luminancia.
2. Dividir cada secuencia HV y HVBAR en regiones de 8 x 8 x 0,2 segundos.
3. Calcular la media de cada una de las regiones.
4. Aplicar un umbral de percepción, reemplazando por 3 los valores menores.
5. Calcular la relación (HV / HVBAR).
6. Comparar las características originales y procesadas usando la función de comparación racional.
7. Reducción espacial calculando la media del 5% de bloques peores para cada 0,2 segundos de tiempo.
8. Reducción temporal tomando la media sobre la secuencia completa.
9. Calcular la raíz cuadrada del parámetro y recortar a un valor mínimo de 0,06.

C. *Parámetro "hv_gain"*

Este parámetro detecta una desviación de bordes diagonales hacia bordes horizontales o verticales (al contrario que el anterior). Esto podría ocurrir si el vídeo procesado contiene, por ejemplo, efecto de bloques. El proceso de cálculo es el siguiente:

1. Realizar los pasos del 1 a 5 del parámetro *hv_loss*.
2. Comparar las características originales y procesadas usando la función de comparación logarítmica.
3. Reducción espacial calculando la media del 5% de bloques peores para cada período de 0,2 segundos.
4. Reducir temporalmente con la media de los períodos en la secuencia completa.

D. *Parámetro "chroma_spread"*

Este parámetro detecta cambios en la dispersión de las muestras de color bidimensionales:

1. Divide los planos de color *Cb* y *Cr* en regiones espacio temporales de 8 píxeles x 8 líneas x 1 fotograma.
2. Calcula la media de cada región. Multiplica la media *Cr* por 1,5 para incrementar la ponderación perceptual de la componente roja en el siguiente paso.
3. Compara las características originales y procesadas usando la distancia euclídea.
4. Reduce espacialmente calculando la desviación típica de los bloques para cada fotograma.
5. Reduce temporalmente ordenando los valores en el tiempo y seleccionando el percentil 10, y luego recortando a un valor mínimo de 0,6.

Los pasos 1 y 2 esencialmente submuestrean los planos *Cb* y *Cr*. Los parámetros *si_loss*, *hv_loss* y *hv_gain* examinan los bordes que contienen suficientes píxeles para ser perceptualmente significativos. Por su parte, *chroma_spread* realiza una integración coherente de las muestras de color sobre un área suficientemente grande como para poder tener impacto perceptual significante.

E. *Parámetro "si_gain"*

Este es el único parámetro de mejora de calidad del modelo. Mide mejoras que puedan resultar de procesos como el de definición de bordes. El parámetro *si_gain* se calcula siguiendo los cinco pasos siguientes:

1. Realizar los pasos 1 a 3 del parámetro *si_loss*.

2. Aplicar un umbral de percepción, reemplazando por 8 los valores menores a él.
3. Comparar las características originales y procesadas usando la función de comparación logarítmica.
4. Reducir espacial y temporalmente calculando la media de todos los bloques y luego recortando a un valor mínimo de 0,004. Estos pasos estiman el valor medio de definición de bordes presente.
5. Fijar a 0,14 todos los valores mayores para prevenir mejoras excesivas de la calidad de más de un tercio de la unidad de calidad tras multiplicar el parámetro por su factor de ponderación correspondiente (modelo general). Un tercio de la unidad de calidad es la mejora máxima observada en los experimentos subjetivos realizados para desarrollar este parámetro. No obstante, en el modelo general la medida VQM se recortará para evitar que el parámetro *si_gain* produzca una estimación de calidad de la secuencia procesada mejor que la original.

F. *Parámetro "ct_ati_gain"*

La percepción de distorsión espacial puede estar afectada por la cantidad de movimiento presente. Asimismo, la percepción de distorsiones temporales puede estar afectada por la cantidad de detalle espacial. Una característica derivada del producto de la información de contraste y de la información temporal puede usarse para considerar parcialmente estas interacciones. El parámetro *ct_ati_gain* se procesa como el producto de una característica de contraste (que mide la cantidad de detalle espacial) y una característica de información temporal (que mide la cantidad de movimiento presente en la región espacio-temporal). Las distorsiones serán más visibles en aquellas regiones que tengan un producto bajo y menos visibles en las regiones con producto alto. Con este parámetro se cuantifican las interacciones espacio-temporales de efectos como ruido, errores de bloques y distorsiones de bordes.

1. Aplicar el filtro de detección de movimiento "valor absoluto de información temporal" (ATI, por sus siglas en inglés) a cada plano de luminancia. ATI es el valor absoluto de la diferencia píxel a píxel entre el fotograma actual y el anterior.
2. Dividir cada secuencia de vídeo en regiones de 4 x 4 x 0,2 segundos.
3. Calcular la desviación típica de cada región
4. Aplicar un umbral de percepción, reemplazando por 3 los valores menores.
5. Repetir los pasos 2, 3 y 4 en la secuencia de luminancia (sin filtrado perceptual) para generar las características de contraste.
6. Multiplicar las características de contraste con el filtrado ATI.

7. Comparar las características originales y procesadas usando la función de comparación racional.
8. Reducir espacialmente calculando la media de cada período de 0,2 segundos.
9. Reducir temporalmente ordenando los valores en el tiempo y seleccionando el percentil 10. Los valores de los parámetros son todos positivos.

G. *Parámetro "chroma_extreme"*

Este parámetro usa las mismas características de color que la medida *chroma_spread*, pero usa unas funciones de reducción distintas. Con él se detectan distorsiones localizadas de color, como las que se producen por errores en transmisiones digitales.

1. Realizar los pasos de 1 a 3 del parámetro *chroma_spread*.
2. Reducir espacialmente calculando para cada período de tiempo la media del 1% de bloques peores y restando a ese resultado el percentil 99. Esto identifica distorsiones graves que afectan a un área pequeña de la imagen.
3. Reducir temporalmente mediante la desviación típica de los resultados del paso 2.

4.1.7. Modelo general

A continuación se describe cómo obtener la medida del modelo general usando los parámetros calculados anteriormente. El modelo general está optimizado para alcanzar la máxima correlación entre resultados subjetivos y objetivos usando videos con un amplio rango de calidades y tasas de bits. Los parámetros descritos cuantifican efectos perceptuales que abarcan una gran gama de distorsiones: desenfoque, distorsión de bloques, vibración en el movimiento, ruido (tanto en el plano de luminancia como en los de crominancia) y errores de bloque (normalmente presentes al ocurrir errores de transmisión digital). Este modelo consiste en una combinación lineal de los parámetros descritos. El resultado varía desde 0 (sin distorsión perceptible) hasta aproximadamente 1 (máxima distorsión perceptible).

El modelo de calidad de vídeo general (VQM) consiste en la siguiente combinación lineal de los siete parámetros explicados anteriormente:

$$
\begin{aligned}
VQM = \; & -0{,}2097 \cdot si\_loss \\
& +0{,}5969 \cdot hv\_loss \\
& +0{,}2483 \cdot hv\_gain \\
& +0{,}0192 \cdot chroma\_spread \\
& -2{,}3416 \cdot si\_gain \\
& +0{,}0431 \cdot ct\_ati\_gain \\
& +0{,}0076 \cdot chroma\_extreme
\end{aligned}
$$

Nótese que *si_loss* es siempre menor o igual a cero, por lo que *si_loss* únicamente puede aumentar la VQM. Al ser el resto de parámetros mayores o iguales a cero, *si_gain* es el único que puede disminuir el valor de VQM.

Después del cálculo, el resultado se recorta a un umbral inferior de 0. Con ello se evita que *si_gain* pueda producir un valor de calidad que valore mejor a la secuencia procesada que a la original (VQM negativo). Por último, se aplica una función de contracción que permite un exceso máximo del 50% a valores mayores de 1,0. Con ello se limita el resultado para videos excesivamente distorsionados que caen fuera del rango de los experimentos subjetivos.

$$\text{Si } VQM > 1{,}0 \quad \text{entonces} \quad VQM = (1 + c) \cdot VQM / (c + VQM)$$

El valor VQM calculado como se indica en la fórmula anterior tendrá valores mayores o iguales a cero y un valor nominal máximo de uno. Sin embargo, el resultado puede exceder ese valor para escenas de vídeo extremadamente distorsionadas.

4.2 Medida de Calidad de Vídeo Digital de Watson modificada

Como se ha comentado anteriormente, la medida DVQ [57] se basa en la medida de Calidad de Vídeo de Watson; que calcula la visibilidad de artefactos haciendo uso de la Transformada Discreta del Coseno (DCT, por sus siglas en inglés).

La medida DVQ es un intento de incorporar numerosos aspectos de la sensibilidad visual humana en un sencillo algoritmo. La sencillez es un objetivo importante, ya que se desearía poder ejecutar la medida en tiempo real sin unos requisitos computacionales

excesivos. Uno de los aspectos más complejos y costosos computacionalmente de este tipo de medidas son las operaciones de filtrado espacial empleadas para implementar el conjunto de filtros paso banda que son característicos de la visión humana. En la medida DVQ se acelera este paso usando la Transformada Discreta del Coseno para la descomposición en canales espaciales. Esto proporciona una importante ventaja ya que existe hardware muy eficiente para esta transformación y porque en muchas aplicaciones la transformación puede haberse realizado ya como parte del proceso de compresión.

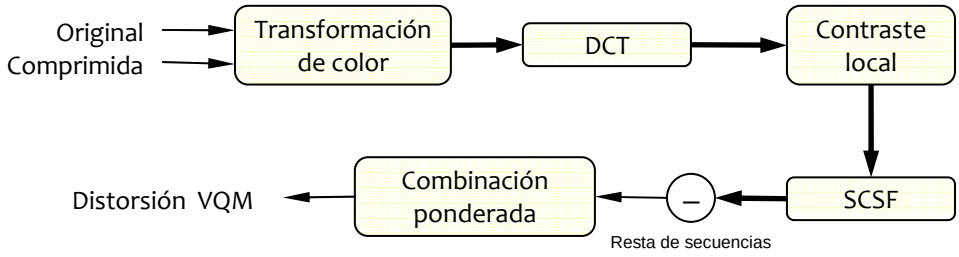


Fig. 4.2 Diagrama de flujo del algoritmo DVQ modificado

La entrada del algoritmo es un par de secuencias a color: la original o referencia y la comprimida o de prueba. Ambas secuencias se convierten al espacio de color YOZ, para después calcular la transformada DCT. Los coeficientes DCT se convierten a unidades locales de contraste, que se definen como la relación de la amplitud de los valores DCT con el valor de continua para el bloque correspondiente. Éstas son procesadas mediante funciones de sensibilidad al contraste espacial para fotogramas estáticos y dinámicos y los coeficientes DCT se convierten a unidades JND. Las secuencias de vídeo son restadas para producir una secuencia diferencia que se procesa mediante una máscara de contraste. Por último se obtiene el valor de distorsión mediante una combinación ponderada.

4.2.1. Entrada

El primer paso consiste en varias transformaciones de muestreo, recorte, y transformación de color que restringen el procesado a una región de interés y expresan las secuencias en un espacio de color perceptual. Esta etapa también comprende el desentrelazado y una corrección gamma de los vídeos de entrada.

El espacio de color YOZ se usa en el modelado de errores perceptuales en compresión de imágenes. Con él la respuesta de los tres tipos de conos del sistema visual humano se pueden transformar en una banda de luminancia (Y) y dos bandas de color opuestas: rojo-verde (O) y azul-amarillo (Z). El espacio de color YOZ deriva del estándar de coordenadas de cromaticidad XYZ desarrollado por la Comisión Internacional de Iluminación (CIE, por sus siglas en francés) con la intención de simular el sistema triestímulo de percepción humana. La transformación XYZ a YOZ es la siguiente:

$$[YOZ] = [XYZ] \cdot \begin{bmatrix} 0 & 0{,}47 & 0 \\ 1 & -0{,}37 & 0 \\ 0 & -0{,}10 & 1 \end{bmatrix}$$

En este punto se aplica la transformada DCT por bloques 8 x 8 en cada canal, separando los fotogramas de entrada en diferentes componentes de frecuencia espacial.

4.2.2. Contraste local

Los coeficientes de la DCT se convierten a unidades de contraste local de la siguiente forma. Primero, se extraen las componentes de continua de todos los bloques. Estos son entonces filtrados en el tiempo, usando un filtro paso bajo IIR de primer orden con una ganancia de 1 y una constante temporal de τ_1. Los coeficientes de la DCT son posteriormente divididos por los coeficientes de continua filtrados de cada bloque correspondiente. Los bloques Z e Y se dividen por los coeficientes de continua Z e Y; el O se divide por la componente de continua del canal Y. En cada caso, una constante muy pequeña se añade al divisor para evitar la división por 0. Finalmente, los cocientes son ajustados a las magnitudes relativas correspondientes a una función basada en unidades de contraste. Se convierte cada coeficiente DCT en un número entre -1 y 1, que expresa la amplitud de la función como una fracción de la luminancia media de ese bloque.

Por su parte, los coeficientes de continua se convierten de un modo similar: el valor medio de continua del fotograma completo se sustrae, y el resultado se divide por la media.

4.2.3. Conversión a JNDs

Este paso es diferente del modelo de Watson. En lugar de aplicar un filtrado temporal y una función de sensibilidad humana al contraste espacial (SCFC, por sus siglas en inglés) de forma separada, se opta por aplicar una matriz SCSF para fotogramas estáticos y otra matriz para fotogramas dinámicos en un paso. Esto reducirá en gran medida la carga computacional y los requisitos de memoria. Los coeficientes de la DCT se convierten a unidades JND multiplicando cada coeficiente por su entrada correspondiente en la matriz SCSF. Para la matriz estática se hace uso de la matriz de cuantificación por defecto de MPEG (que puede ser vista como la matriz de umbral al contraste, inversa de la matriz de sensibilidad al contraste), ya que esta matriz está basada en una amplia investigación psicofísica. Para la matriz dinámica, se eleva cada entrada en la matriz SCSF estática a una potencia para tener en cuenta la propiedad temporal de la SCSF. La potencia se decide según la tasa de fotogramas de las secuencias de vídeo.

$$\begin{bmatrix} 8 & 16 & 19 & 22 & 26 & 27 & 29 & 34 \\ 16 & 16 & 22 & 24 & 27 & 29 & 34 & 37 \\ 19 & 22 & 26 & 27 & 29 & 34 & 34 & 38 \\ 22 & 22 & 26 & 27 & 29 & 34 & 37 & 40 \\ 22 & 26 & 27 & 29 & 32 & 35 & 40 & 48 \\ 26 & 27 & 29 & 32 & 35 & 40 & 48 & 58 \\ 26 & 27 & 29 & 34 & 38 & 46 & 56 & 69 \\ 27 & 29 & 35 & 38 & 46 & 56 & 69 & 83 \end{bmatrix}$$

Fig. 4.3 Matriz de cuantificación por defecto de MPEG

4.2.4. Combinación ponderada de distorsión media y máxima

Previamente, las dos secuencias son restadas. En este paso VQM también difiere de DVQ al incorporar enmascaramiento de contraste consistente en una simple operación de máximo que luego pondera junto con la distorsión media. Esto refleja el hecho de que una gran distorsión en una región suprimirá nuestra sensibilidad a otra distorsión menor.

$$Mean\_dist = 1000 \cdot mean\ (mean\ (abs\ (diff)\ )\ )$$
$$Max\_dist = 1000 \cdot maximum\ (maximum\ (abs\ (diff)\ )\ )$$

$$VQM = (\ Mean\_dist + 0{,}005 \cdot Max\_dist\ )$$

El parámetro de ponderación de la distorsión máxima 0,005 se elige basándose en numerosos experimentos psicofísicos previos. El parámetro 1000 es la relación de estandarización.

4.3 Medida de Calidad de Vídeo VSSIM

Las señales de imágenes naturales están altamente estructuradas. Por "señal estructurada" se entiende que las muestras de la señal tienen una alta dependencia entre ellas, especialmente si están próximas en el espacio. Sin embargo, como se vio al estudiar medidas de distorsión estructural, la fórmula de combinación de Minkowski usada en los modelos basados en sensibilidad al error es una operación de diferenciación punto a punto independiente de la estructura de la señal. Además, la descomposición de la señal usando transformaciones lineales, puede no eliminar completamente las dependencias entre las muestras de la señal. Incluso un número significativo de cambios estructurales en la señal pueden no ser capturados con la fórmula de Minkowski.

El objetivo de esta nueva medida es encontrar una forma más directa de comparar las estructuras de las secuencias original y distorsionada. La principal diferencia con respecto a los algoritmos basados en sensibilidad al error es que las degradaciones de la imagen se consideran pérdidas en la información estructural percibida en lugar de errores percibidos.

4.3.1. Índice de Similitud Estructural (SSIM)

Existen numerosas interpretaciones de la nueva filosofía, dependiendo de cómo se interpreten y cuantifiquen los conceptos de "información estructural" y "distorsión estructural". Aquí, desde un punto de vista de formación de la imagen, se considerará "información estructural" de la imagen todos aquellos atributos que reflejen la estructura de los objetos de la escena, independientemente de la luminancia media y del contraste de la imagen. Esto conduce a una aproximación de medida de calidad que separe la luminancia, el contraste y las distorsiones estructurales.

A continuación se explica un algoritmo de indexación de imágenes por similitud [58]. Sean x e y dos señales no negativas que han sido alineadas entre ellas, y sean μ_x, μ_y, σ_x^2, σ_y^2 y σ_{xy} la media de x, la media de y, la varianza de x, la varianza de y, y la covarianza de x e y, respectivamente. La media y la desviación típica (raíz cuadrada de la varianza) de una señal se consideran burdas estimaciones de la luminancia de la señal. La covarianza (normalizada por la varianza) puede considerarse como una medida de cuánto cambio no lineal posee una señal con respecto a la otra. Se definen las medidas de luminancia, contraste y comparación estructural de la siguiente forma:

$$l(x,y)=\frac{2\mu_x\mu_y}{\mu_x^2+\mu_y^2},\quad c(x,y)=\frac{2\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2},\qquad s(x,y)=\frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

Es importante destacar que estos cálculos son conceptualmente independientes en el sentido de que los dos primeros valores sólo dependen de la luminancia y el contraste de las imágenes a comparar, respectivamente, y un cambio únicamente en la luminancia o el contraste de las imágenes no tiene impacto en el tercer valor. Geométricamente, *s(x,y)* corresponde al coseno del ángulo entre los vectores x- μ_x e y- μ_y, independientemente de la longitud de esos vectores. Aunque *s(x,y)* no usa una descripción representativa directa de las estructuras de las imágenes, refleja la similitud entre las dos estructuras – es igual si y sólo si las estructuras de las dos imágenes en comparación son exactamente la misma (recordemos que se considera información estructural todos aquellos atributos que no son la información de luminancia ni de contraste).

Cuando $\left(\mu_x^2+\mu_y^2\right)\cdot\left(\sigma_x^2+\sigma_y^2\right)\neq 0$, el índice de similitud entre x e y corresponde a:

$$S(x,y)=l(x,y)\cdot c(x,y)\cdot s(x,y)=\frac{4\mu_x\mu_y\sigma_{xy}}{\left(\mu_x^2+\mu_y^2\right)\cdot\left(\sigma_x^2+\sigma_y^2\right)}$$

Si las dos señales se representan de forma discreta como $x=\{x_i|i=1,2,\cdots,N\}$ e $y=\{y_i|i=1,2,\cdots,N\}$, entonces los parámetros estadísticos pueden ser estimados de la siguiente forma:

$$\mu(x) = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \mu(y) = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad \sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})\cdot(y_i - \bar{y})$$

Un problema con el cálculo de *S(x,y)* es que si $(\mu_x^2 + \mu_y^2)\cdot(\sigma_x^2 + \sigma_y^2)$ está próximo a 0, la medida resultante se vuelve inestable. Este efecto se ha observado frecuentemente en los experimentos, especialmente es áreas lisas de la imagen. Para evitar este problema, se ha modificado la ecuación de *S(x,y)*. La nueva medida resultante es el Índice de Similitud Estructural (SSIM, *Structural SIMiliarity index*) entre las señales *x* e *y*:

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

Se añaden dos constantes, C_1 y C_2:

$$C_1 = (K_1 \cdot L)^2 \quad \text{y} \quad C_2 = (K_2 \cdot L)^2$$

donde *L* es el rango dinámico de los valores de los píxeles (para 8bits/píxel de imágenes en escala de gris, *L*=255), y K_1 y K_2 son dos constantes cuyos valores han de ser lo suficientemente pequeños de forma que C_1 y C_2 sólo tengan efecto cuando $(\mu_x^2 + \mu_y^2)$ o $(\sigma_x^2 + \sigma_y^2)$ sean pequeños.

A partir de los experimentos realizados en la prueba de este índice, se fijan K_1=0,01 y K_2=0,03. El índice SSIM satisface las siguientes condiciones:

SSIM(x,y) = SSIM(y,x);
SSIM(x,y) ≤ *1*;
SSIM(x,y) = 1 si y sólo si *x=y* (en representación discreta, $x_i = y_i$ para todo *i*=1,2,⋯,N);

Basado en la filosofía descrita anteriormente, si consideramos una de las imágenes a comparar de calidad perfecta, entonces el índice SSIM proporcionará una medida cuantitativa de la calidad de la otra imagen.

El algoritmo SSIM se aplica en evaluación de calidad de imágenes estáticas usando una aproximación de ventana deslizante. El tamaño de la ventana se fija a 8x8 píxeles. El índice SSIM se calcula para la región contenida en la ventana, que se desplaza píxel a píxel desde la parte superior izquierda a la parte inferior derecha de la imagen. Esto resulta en un mapa de índices SSIM de la imagen, que se considera el mapa de calidad de la imagen en evaluación. El valor global se define como la media del mapa de calidad, es decir, el índice SSIM medio (MSSIM).

4.3.2. Medida de calidad de vídeo (VSSIM)

En este apartado se explicará un método que emplea el índice SSIM como medida única para varios tipos de distorsiones de secuencias de vídeo. A continuación se muestra el diagrama del sistema de evaluación de calidad de vídeo propuesto. La calidad del vídeo distorsionado se mide en tres niveles: el nivel de área local, el nivel del fotograma completo y el nivel de secuencia.

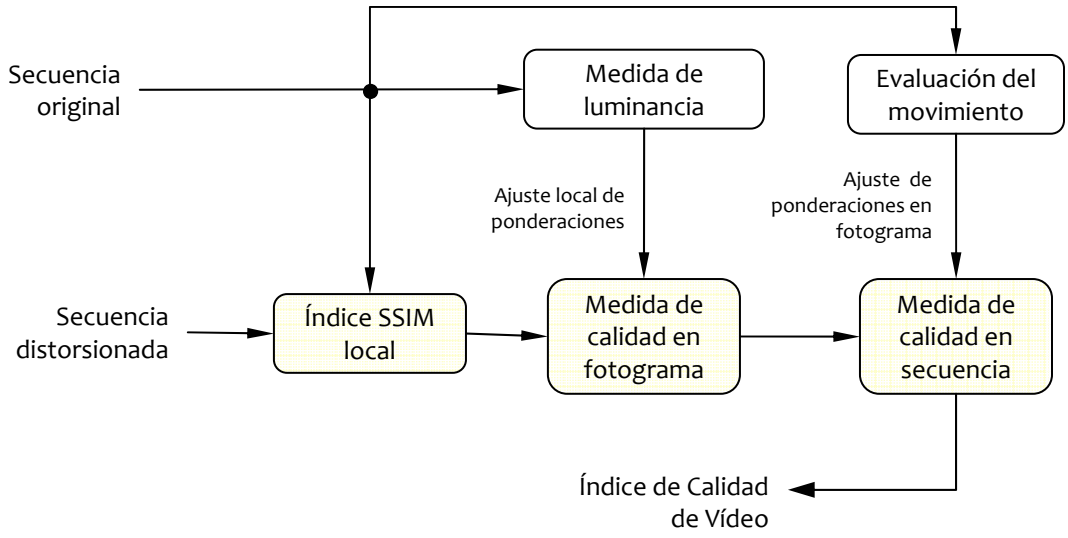


Fig. 4.4 Diagrama de flujo del algoritmo VSSIM

En primer lugar, se extraen áreas de muestreo local del fotograma correspondiente en una localización espacial determinada tanto en la secuencia original como en la distorsionada. Las zonas de muestreo son seleccionadas de forma aleatoria en ventanas de 8x8 píxeles. Esto difiere de la medida de imágenes estáticas vista anteriormente donde todas las posibles ventanas de muestreo eran seleccionadas mediante una ventana deslizante que se movía píxel por píxel a través de la imagen completa. En lugar de ello, se seleccionan únicamente una parte de todas las ventanas 8x8 posibles. Se emplea el número de ventanas de muestreo por fotograma (*Rs*) para representar la densidad de muestreo. Los experimentos demuestran que un parámetro *Rs* seleccionado apropiadamente puede reducir el coste computacional de forma muy significativa mientras que se mantienen resultados razonablemente robustos.

El índice SSIM se aplica entonces a las componentes Y, Cb y Cr de forma independiente y son combinados en una medida de calidad local usando una suma ponderada.

Sean $SSIM_{ij}^{Y}$, $SSIM_{ij}^{Cb}$ y $SSIM_{ij}^{Cr}$ los índices SSIM para las componentes *Y*, *Cb* y *Cr* de la ventana de muestreo *j-ésima* en el fotograma *i-ésimo*. El índice de calidad local resulta:

$$SSIM_{ij} = W_Y \cdot SSIM_{ij}^{Y} + W_{Cb} \cdot SSIM_{ij}^{Cb} + W_{Cr} \cdot SSIM_{ij}^{Cr}$$

Donde las ponderaciones se fijan según las pruebas experimentales realizadas como:

$$W_Y = 0{,}8; \qquad W_{Cb} = 0{,}1; \qquad W_{Cr} = 0{,}1;$$

En el segundo nivel de evaluación de la calidad, los valores de calidad local se combinan en un índice a nivel de fotograma usando la siguiente fórmula:

$$Q_i = \frac{\sum_{j=1}^{Rs} w_{ij} \cdot SSIM_{ij}}{\sum_{j=1}^{Rs} w_{ij}}$$

donde Q_i denota la medida del índice de calidad del fotograma *i-ésimo* en la secuencia de vídeo, y w_{ij} es el valor de ponderación dado a la ventana *j-ésima* en el fotograma *i-ésimo*.

Finalmente, en el tercer nivel la calidad global de la secuencia de vídeo completa viene dada por:

$$Q = \frac{\sum_{j=1}^{F} W_i \cdot Q_i}{\sum_{j=1}^{F} W_i}$$

donde F es el número de fotogramas y W_i es el valor de ponderación asignado al fotograma *i-ésimo*.

Si todos los fotogramas y todas las ventanas de muestreo en cada fotograma se consideran de igual modo, entonces:

$wij = 1$ para todo i,j y $\sum_{j=1}^{Rs} w_{ij} = R_s$ para todo i

Esto da lugar a una medida de calidad equivalente al índice SSIM medio de todas las ventanas de muestreo en todos los fotogramas. Estos valores de ponderación no serán óptimos ya que diferentes regiones y diferentes fotogramas tendrán distinta importancia para los observadores humanos. La asignación óptima de valores de ponderación es difícil ya que influyen numerosos aspectos psicológicos, que pueden depender del contenido y del contexto de la secuencia en observación. Sin embargo, ciertos ajustes al modelo de pesos iguales pueden ayudar a mejorar la precisión de la estimación del algoritmo.

Se emplearán dos simples métodos de ajuste. El primero se basa en la consideración de que las zonas oscuras normalmente no atraen la atención del observador, por lo que se deberían ponderar con pesos menores. Usamos el valor medio μ_x de la componente Y como un estimador de la luminancia local, y los valores de ponderaciones locales se ajustan como sigue:

$$wi_j = \begin{cases} 0 & \mu_x \le 40 \\ (\mu_x - 40)/10 & 40 < \mu_x \le 50 \\ 1 & \mu_x > 50 \end{cases}$$

El segundo ajuste considera el caso en el que tiene lugar una gran cantidad de movimiento global. En efecto, algunas distorsiones de imágenes se perciben de forma diferente cuando el fondo del vídeo se está moviendo de forma muy rápida (normalmente corresponde a rápidos movimientos de cámara). Por ejemplo, un desenfoque muy alto se percibe normalmente como una distorsión muy desagradable en imágenes fijas o en vídeo de movimiento muy lento. Sin embargo, la misma cantidad de desenfoque puede no ser tan importante en una secuencia con mucho movimiento, porque al mismo tiempo tiene lugar un importante desenfoque perceptual de movimiento. Ese tipo de diferencias no pueden ser capturadas por el índice SSIM intra-fotograma, que no considera ninguna información de movimiento. Las pruebas experimentales también han demostrado que el algoritmo se comporta de manera más inestable cuando gran cantidad de movimiento global tiene lugar. Por lo tanto, se da una ponderación menor a los fotogramas en los que existe mucho movimiento para mejorar la robustez del algoritmo. En primer lugar, para cada ventana de muestreo, se usa un algoritmo de estimación de movimiento basado en bloques para evaluar su movimiento con respecto al fotograma siguiente. Supongamos que m_{ij} representa la longitud del vector de movimiento de la ventana de muestreo *j-ésima* en el fotograma *i-ésimo*, entonces el nivel de movimiento del fotograma *i-ésimo* se estima como:

$$M_i = \frac{\left(\sum_{j=1}^{Rs} m_{ij} \right) / R_s}{K_M}$$

donde K_M es una constante que sirve como factor de normalización del nivel de movimiento del fotograma. El valor que se usa en los experimentos es K_M = 16. La ponderación de cada fotograma se ajusta entonces como:

$$W_i = \begin{cases} \sum_{j=1}^{Rs} w_{ij} & M_i \le 0{,}8 \\ \left((1{,}2 - M_i)/0{,}4\right) \sum_{j=1}^{Rs} w_{ij} & 0{,}8 < M_i \le 1{,}2 \\ 0 & M_i > 1{,}2 \end{cases}$$

Sin embargo, éste es sólo uno de los ajustes posibles al modelo de pesos iguales. En el siguiente capítulo se describirá el método de ajuste empleado en la medida propuesta.

5

Medida de Calidad de Vídeo propuesta

Una vez analizado el problema de la evaluación objetiva de la calidad perceptual de vídeo y tras estudiar algunos de los algoritmos más extendidos actualmente, en este capítulo se describirá la métrica propuesta. Se detallarán las técnicas en las que se basa, su estructura, las decisiones de diseño e implementación, así como los bloques adicionales presentes en la medida. También se realizará una descripción general de la aplicación desarrollada, indicando detalles de la implementación, parámetros de entrada – salida, y arquitectura general del sistema.

Los requisitos de diseño del algoritmo se centraban en lograr una medida sin excesiva carga computacional que obtuviera unos resultados razonablemente buenos. Comparando las técnicas estudiadas, se puede observar una clara diferencia entre dos de ellas (la eficacia de cada una de las medidas descritas se analizará con detalle en el capítulo 6, en la comparación de algoritmos). Por un lado, la medida de Watson modificada proporciona un tiempo de ejecución realmente bajo, pero con unos resultados no muy precisos. Por el contrario, la medida NTIA ofrece unos resultados muy buenos, pero a costa de una carga computacional muy elevada que provoca un tiempo de ejecución bastante importante.

El objetivo principal, por lo tanto, es desarrollar una métrica que mejore los resultados de las actuales medidas con tiempo de ejecución bajo, sin aumentar excesivamente la carga computacional, con el fin de poder ser implantada en sistemas de tiempo real. Para ello se

opta por las nuevas técnicas basadas en distorsión estructural, que se han mostrado más precisas que las basadas en detección de error.

5.1 Estructura del algoritmo

La medida de calidad de vídeo propuesta se basa principalmente en el Índice de Similitud Estructural de Vídeo (VSSIM), incorporando algunas modificaciones; sin embargo, adicionalmente se hace uso de la reciente medida de predicción de MOS (MOSp) desarrollada por A. Bhat, I. Richadson y S. Kannangara [59]. Se opta por esta última medida debido a la buena correlación obtenida con el MOS, sobre todo en secuencias con baja actividad espacial y temporal, donde toma un peso más importante en el índice de calidad final.

A modo de introducción muy general, se muestra a continuación un diagrama que ilustra la arquitectura global de la medida propuesta. La figura 5.7 proporciona una descripción más completa de la estructura.

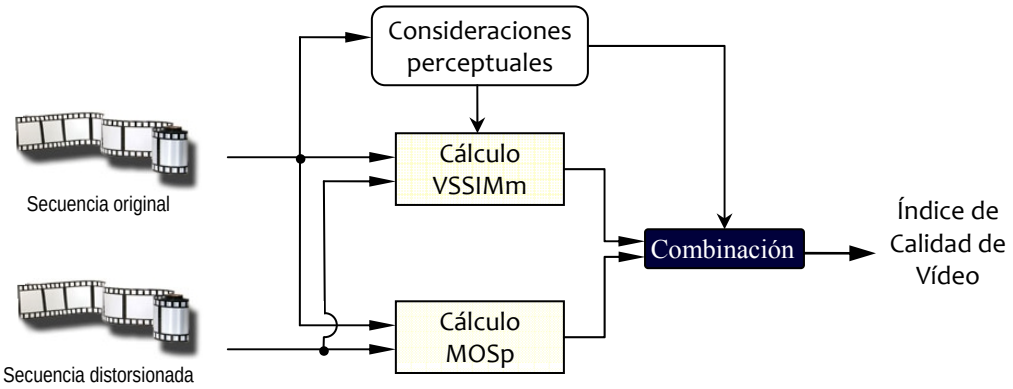


Fig. 5.1 Esquema global de la estructura de la medida

En los siguientes apartados se describen las dos técnicas utilizadas; posteriormente se indicará cómo se realiza la integración de ambas en el sistema para generar el índice final de calidad de vídeo.

5.1.1 VSSIM modificado

En el caso del VSSIM ya se ha realizado una descripción previa en el apartado 4.3. Se opta por esta medida por tratarse de una técnica basada en distorsión estructural que, aún sin ofrecer unos resultados excesivamente buenos, cuenta con grandes posibilidades de ampliación. Como se ha visto, el VSSIM se basa en el índice de similitud estructural (SSIM) de imágenes. Añadiendo nuevos bloques y algoritmos se consigue extender esa información al dominio temporal, con lo que resulta una medida de calidad de vídeo cuyos resultados dependerán de la eficacia de los nuevos algoritmos añadidos.

Como recordatorio de la estructura del algoritmo SSIM, se muestran los pasos necesarios para la obtención del índice.

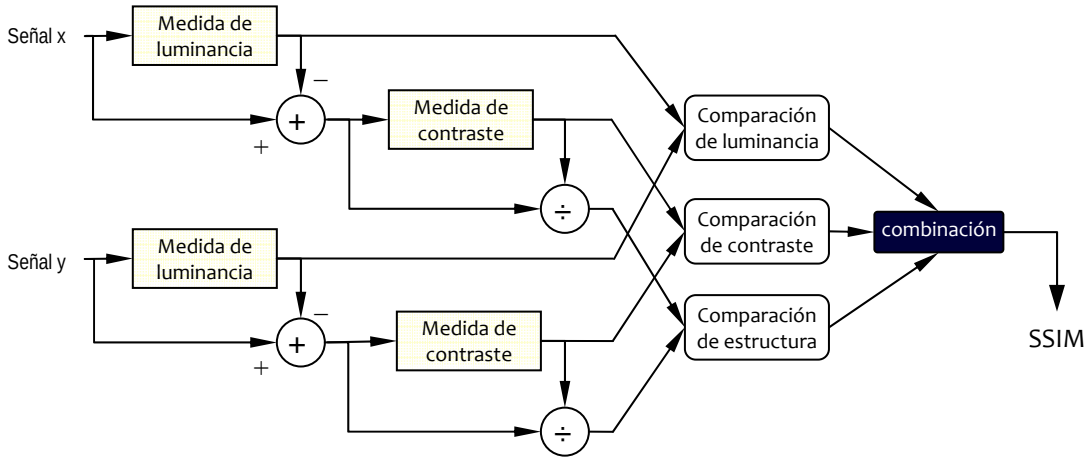


Fig. 5.2 Diagrama del sistema de medida de similitud estructural (SSIM)

Partiendo del cálculo del SSIM y haciendo uso de las técnicas del VSSIM, a continuación se detallan los cambios y ajustes realizados a la medida para la obtención del índice modificado (VSSIMm) que forma parte del cálculo final.

En primer lugar se seleccionan zonas aleatorias de muestreo en bloques de 16x16 píxeles. El tamaño de ventana difiere con el propuesto por los creadores de la medida (8x8 píxeles); se elige este valor para recoger de forma más amplia la estructura de la imagen, así

como por compatibilidad con algunos de los bloques adicionales (como por ejemplo la ponderación por luminancia) que se explicarán más adelante.

Se procesan únicamente una parte de todas las ventanas 16x16 posibles. Tras numerosas pruebas experimentales con el fin de determinar el número de ventanas de muestreo por fotograma, el parámetro *Rs* (que representa la densidad de muestreo) se fija por defecto a 0,1. No obstante, el usuario puede variar este valor en función del compromiso precisión – tiempo de ejecución. Aunque en un principio el valor elegido pueda parecer bajo, los experimentos demuestran que los resultados se mantienen razonablemente robustos reduciendo de forma muy importante la carga computacional.

Rs	0,1	0,5	1
Tiempo de ejecución	19 s.	143 s.	549 s.
Variabilidad en el resultado	8,3 %	2,4 %	0 %

Tabla 5.1 Comparativa precisión – tiempo de ejecución para una secuencia CIF de 6 s. de duración

Por lo tanto, para cada uno de los bloques se calcula su Índice de Similitud Estructural. Según lo descrito en el capítulo anterior:

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

siendo

$$C_1=(K_1\cdot L)^2 \quad \text{y} \quad C_2=(K_2\cdot L)^2$$

con unos valores de

L = 255 (rango dinámico de los píxeles)
K_1=0,01;
K_2=0,03;

El índice de cada bloque es una combinación ponderada de los índices de las tres componentes *Y*, *Cb* y *Cr*:

$$SSIM(i,j)=W_Y\cdot SSIM_Y(i,j)+W_{Cb}\cdot SSIM_{Cb}(i,j)+W_{Cr}\cdot SSIM_{Cr}(i,j)$$

Con unos factores de ponderación:

$W_Y = 0,8;$
$W_{Cb} = 0,1;$
$W_{Cr} = 0,1;$

Posteriormente se combinan los índices de todos los bloques de un fotograma mediante:

$$Q(i) = \frac{\sum_{j=1}^{Rs} w(i,j) \cdot SSIM(i,j)}{\sum_{j=1}^{Rs} w(i,j)}$$

donde *Q(i)* denota la medida del índice de calidad del fotograma *i-ésimo* en la secuencia de vídeo, y *w(i,j)* es el valor de ponderación dado a la ventana *j-ésima* en el fotograma *i-ésimo*.

Finalmente, en el tercer nivel la calidad global de la secuencia de vídeo completa viene dada por:

$$Q = \frac{\sum_{j=1}^{F} W(i) \cdot Q(i)}{\sum_{j=1}^{F} W(i)}$$

donde *F* es el número de fotogramas y *W(i)* es el valor de ponderación asignado al fotograma *i-ésimo*. Con ello se obtiene el valor Q, o índice de calidad VSSIM modificado.

En el apartado 5.2 de este capítulo se analizarán los valores asignados a los factores de ponderación tanto a nivel de fotograma como a nivel de secuencia; en función de los distintos módulos con consideraciones perceptuales añadidos a la medida.

5.1.2 MOSp

Esta medida perceptual se basa en el uso del MSE; su diseño está motivado por:

a) la obtención de una buena correlación con el MOS
b) mantener una carga computacional baja

Para ello, los autores de la medida estudiaron la teoría de que la correlación entre MSE y MOS tiende a ser alta para una única secuencia codificada a distintas tasas. Durante la investigación de la medida, emplearon ocho secuencias CIF de 10 segundos de duración y codificadas usando el estándar de compresión H264. Las secuencias utilizadas fueron

Carphone, *Foreman*, *Deadline*, *Tempete*, *News*, *Bus*, *Paris* y *Akiyo*, comprimidas con unos valores de *QP* = {6, 26, 34, 36, 38, 40, 42, 45}. Las pruebas subjetivas fueron realizadas con 30 sujetos sin experiencia y siguiendo la recomendación ITU-BT.500.

Fig. 5.3 Secuencias empleadas en el diseño de la medida MOSp

Al observar la relación entre MSE y MOS mediante la figura 5.4 (extraída de [59]) se comprobó la característica para las cuatro escenas de prueba: *Carphone*, *News*, *Paris* y *Bus*. Las curvas son aproximadamente lineales desde MOS=1,0 hasta MOS=0,1 con una variación en la pendiente por debajo de 0,1.

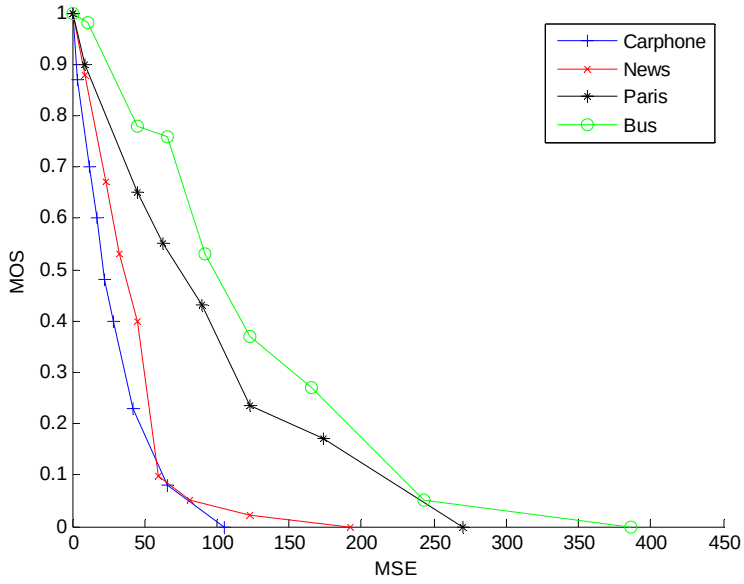


Fig. 5.4 Gráfico de la relación MSE - MOS

Por lo tanto, se realiza un corte del MOS en 0,1 y se trabaja con los datos restantes. Basándose en la relación, la medida perceptual propuesta por los autores fue:

$$MOSp = 1 - k_s(MSE)$$

Con ella se pretende estimar (predecir) la puntuación media de opinión (MOSp) de una secuencia comprimida, empleando el error cuadrático medio (MSE) entre los vídeos original y comprimido y la pendiente de la recta de ajuste (k_s), que se calcula de forma automática a partir del contenido de la secuencia.

La figura 5.5 a. (extraída de [59]) ilustra el modelo propuesto representando la relación lineal entre MOS y MSE donde la máxima calidad percibida (MOS = 1) se observa cuando no hay errores en los píxeles. La figura 5.5 b. (extraída de [59]) muestra el modelo propuesto (líneas en discontinua) de ajuste a las cuatro secuencias.

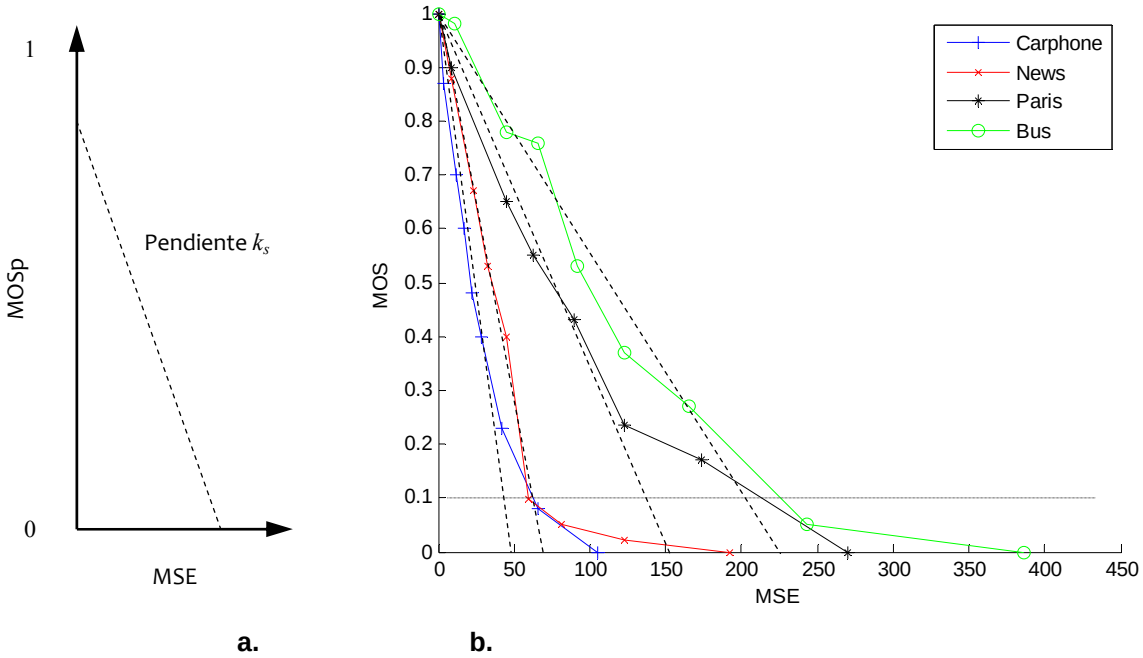


Fig. 5.5 **a.** Modelo propuesto **b.** Rectas de ajuste propuestas para las cuatro secuencias

Hay una clara diferencia en la pendiente de cada secuencia. La medida MOSp propone calcular dicha pendiente a partir de la actividad de la secuencia, que se basa en el enmascaramiento espacial y temporal del vídeo original. Como se ha visto, el fenómeno de enmascaramiento es la causa de que artefactos similares sean más visibles en algunas zonas

determinadas que en otras. Concretamente, el enmascaramiento espacial se da en regiones de la escena con texturas complejas. Los bordes proporcionan una buena estimación de la textura, por lo que se usó la potencia de bordes como medida de información de textura. Para la medida, optaron por una detección de bordes mediante filtros Sobel (debido a su simplicidad y eficiencia) en la componente de luminancia de los fotogramas originales. Por lo tanto, se obtienen los bordes horizontales y los bordes verticales de las imágenes, y el valor total se calcula como:

$$G(x,y) = |G_{horizontal}(x,y)| + |G_{vertical}(x,y)|$$

donde G es la magnitud de borde y *(x,y)* es la localización del píxel. La potencia de los bordes se calcula usando regiones locales de 16x16 píxeles, y la información espacial de textura ($STI_{macrobloque}$) se obtiene como el valor medio de todos los píxeles del macrobloque.

El filtro de Sobel tanto horizontal como vertical se implementa convolucionando una matriz de 3x3 con los macrobloques del fotograma. Siendo $m(x,y)$ el píxel x e y del macrobloque, el resultado de la primera convolución, $G_{horizontal}(x,y)$ se calcula como:

$$\begin{aligned} G_{horizontal}(x,y) = {} & -1 \cdot m(x\text{-}1,y\text{-}1) \; +0 \cdot m(x\text{-}1,y) \; +1 \cdot m(x\text{-}1,y+1) \\ & -2 \cdot m(x,y\text{-}1) \; +0 \cdot m(x,y) \; +2 \cdot m(x,y+1) \\ & -1 \cdot m(x+1,y\text{-}1) \; +0 \cdot m(x+1,y) \; +1 \cdot m(x+1,y+1) \end{aligned}$$

De forma análoga, el resultado de la segunda convolución, $G_{vertical}(x,y)$

$$\begin{aligned} G_{vertical}(x,y) = {} & -1 \cdot m(x\text{-}1,y\text{-}1) \; -2 \cdot m(x\text{-}1,y) \; -1 \cdot m(x\text{-}1,y+1) \\ & +0 \cdot m(x,y\text{-}1) \; +0 \cdot m(x,y) \; +0 \cdot m(x,y+1) \\ & +1 \cdot m(x+1,y\text{-}1) \; +2 \cdot m(x+1,y) \; +1 \cdot m(x+1,y+1) \end{aligned}$$

Los cálculos se realizan para todo $2 \leq x \leq 15$ y $2 \leq y \leq 15$, al ser las regiones locales de tamaño 16 x 16 píxeles.

Tras la obtención de la actividad espacial, se analiza el enmascaramiento temporal, que se produce en regiones con alto nivel de movimiento, pudiendo enmascarar artefactos de forma más efectiva que otras regiones con movimiento menor. De este modo, se empleó

la potencia del gradiente temporal como medida de cambios temporales. El gradiente se calcula como la diferencia absoluta entre el plano de luminancia del fotograma actual ($Y_{fotograma\_i}$) y el plano de luminancia del fotograma previo ($Y_{fotograma\_i-1}$):

$$Y_{dif} = \left| Y_{fotograma\_i} - Y_{fotograma\_i-1} \right|$$

$$TI_{fotograma\_i} = \left| GT_{horizontal}(x,y) \right| + \left| GT_{vertical}(x,y) \right|$$

donde $TI_{fotograma\_i}$ es el gradiente temporal del fotograma actual, $GT_{horizontal}(x,y)$ y $GT_{vertical}(x,y)$ son los gradientes de Sobel horizontal y vertical de la imagen diferencia Y_{dif}. De este modo, un cambio temporal importante entre los píxeles del fotograma actual y del previo resultará en un valor absoluto de diferencia alto y, por tanto, en una magnitud de gradiente elevada. La información temporal de cada macrobloque ($TI_{macrobloque}$) se calcula como la media de todos los píxeles del macrobloque diferencia. La actividad del macrobloque ($Actividad_{macrobloque}$) se obtiene tanto de la información de textura espacial como de la información temporal del macrobloque de la siguiente manera:

$$Actividad_{macrobloque} = max(STI_{macrobloque} , TI_{macrobloque})$$

La actividad de un fotograma se calcula como la media de las actividades de todos los macrobloques del fotograma y la actividad de la secuencia es el valor medio de de todas las actividades de los fotogramas.

Volviendo al modelo, la relación entre la pendiente y la actividad de la secuencia se obtuvo a partir de las ocho secuencias de entrenamiento mencionadas anteriormente. Usando un ajuste exponencial, la relación resultó:

$$k_s = 0{,}03697 \cdot e^{-0{,}02236 \cdot Actividad\_secuencia}$$

Esta curva de ajuste se muestra en la figura 5.6 (extraída de [59]). Se puede comprobar que la relación anterior es una predicción razonable de la pendiente k_s. Además, se observa como en secuencias de baja actividad, un pequeño cambio en el MSE se traduce en un cambio importante en el MOS, mientras que en secuencias de mayor actividad para el mismo cambio de MSE, la variación de MOS es menor.

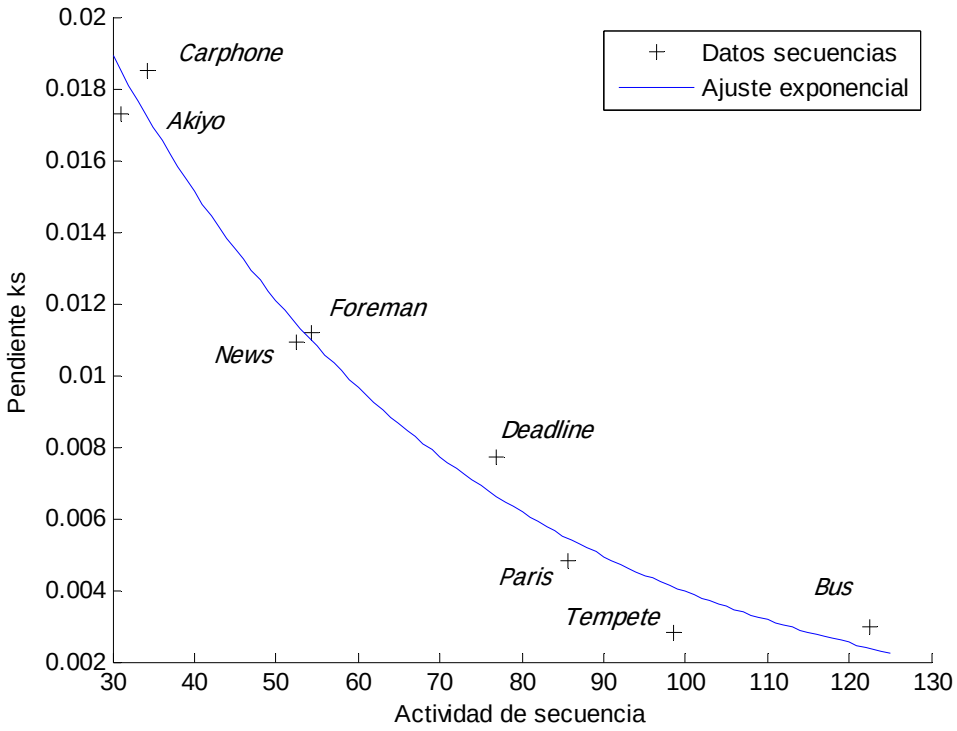


Fig. 5.6 Relación entre actividad de secuencia y pendiente

Una vez obtenido el valor de pendiente del modelo, se calcula la calidad global de la secuencia. Los vídeos distorsionados pueden tener una calidad alta en ciertas zonas de la escena mientras que otras pueden ser de calidad inferior. En estos casos, la calidad de la secuencia será la calidad media. Por lo tanto, la medida evalúa la calidad a nivel de macrobloque en primer lugar (predicción de calidad subjetiva, MOSp). Se calcula la actividad de cada macrobloque con el fin de determinar la pendiente $k_{macrobloque}$. Se obtiene el plano de luminanacia y se calcula el MSE entre los macrobloques originales y los distorsionados. El MOSp para cada macrobloque se obtiene mediante:

$$MOSp_{macrobloque} = 1 - k_{macrobloque}\,(MSE_{macrobloque})$$

Posteriormente se combinan los valores en un índice de calidad a nivel de fotograma. La calidad global del vídeo se obtiene mediante la media de los valores MOSp de todos los fotogramas de la secuencia.

5.1.3 Arquitectura general

Una vez descritas en los apartados anteriores las dos técnicas empleadas en la medida propuesta, así como sus modificaciones, a continuación se indica la estructura general del algoritmo. En la figura inferior se muestra con detalle los pasos necesarios para la obtención de la medida de calidad de vídeo.

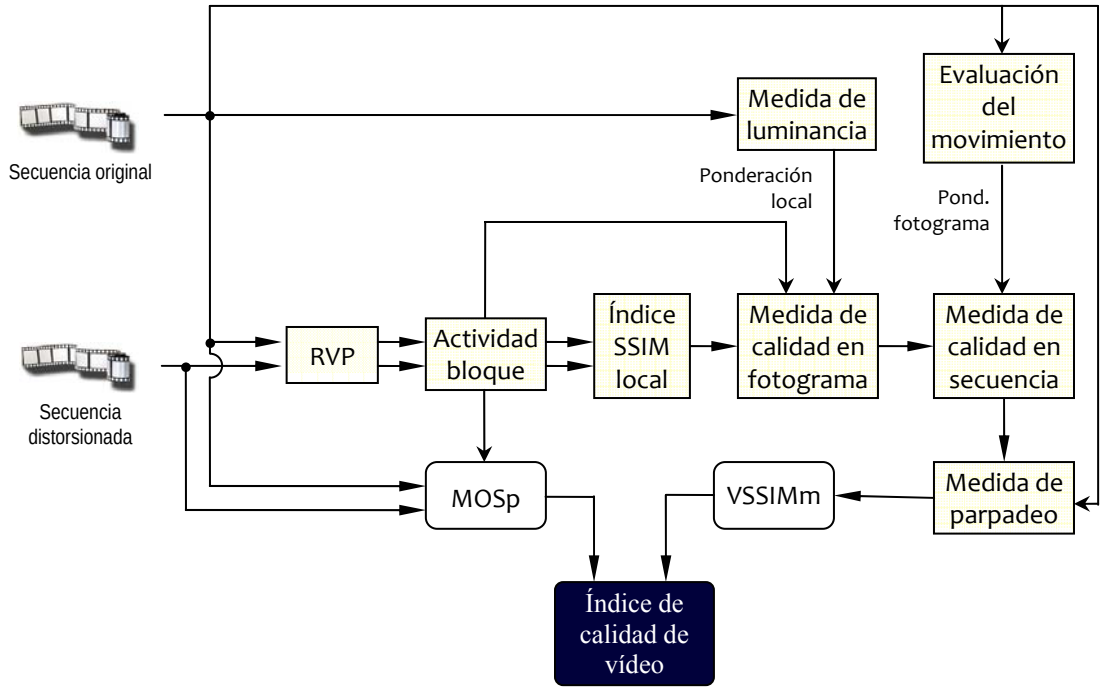


Fig. 5.7 Diagrama detallado de la estructura del algoritmo propuesto

En primer lugar las dos secuencias son procesadas por el módulo RVP, donde se obtiene la Región Válida de Procesado. Tras ello los fotogramas se dividen en regiones de 16x16 píxeles y se calcula la actividad del número de bloques indicado por el parámetro *Rs*. La información de actividad de cada bloque será usada posteriormente para la obtención del MOSp y para la ponderación local de VSSIM. Una vez hallada la actividad, se obtiene el índice SSIM local. Los índices de los bloques de un mismo fotograma se combinan para producir un único valor, para ello se emplean coeficientes de ponderación obtenidos

mediante un módulo de evaluación de luminancia de la secuencia original. Los índices de fotograma se combinan de nuevo para producir un único valor final que represente el VSSIM de la secuencia. En este caso se emplea una ponderación dependiente del nivel de movimiento existente en el vídeo. Por último, dicho valor puede sufrir modificaciones en función del módulo de detección de parpadeo, que lleva a cabo una medida de dicho efecto en la secuencia original.

Por otro lado, para el cálculo del MOSp se procesan ambas secuencias (original y distorsionada) empleando la actividad de bloque calculada anteriormente. Una vez se tienen los valores VSSIM modificado y MOSp, se calcula el índice de calidad de vídeo final. Para ello, se realiza una ponderación en la que, partiendo de una situación proporcional, los pesos se ajustan dinámicamente en función del contenido de la secuencia. En secuencias con un nivel bajo de actividad, tomará una importancia mayor el valor MOSp. A medida que se detecten niveles altos de actividad espacial, temporal o presencia de parpadeo, el valor del VSSIM modificado irá tomando más importancia en el índice final.

En el apartado siguiente se explica con detalle cada uno de los módulos adicionales presentes en el algoritmo, indicando su función, las técnicas empleadas y los resultados que produce.

5.2 Módulos adicionales

Antes de comenzar con la descripción de las funcionalidades añadidas a los algoritmos base para obtener el índice final, es necesario indicar una comprobación previa que realiza el algoritmo. No se trata de un módulo como tal, pero se evitan inestabilidades y proporciona robustez. La verificación consiste en el análisis de los parámetros de ambas secuencias. En concreto se comprueba que sean de la misma duración y que tengan la misma altura y la misma anchura. Si alguno de estos supuestos no se cumple, el algoritmo no continúa su ejecución.

5.2.1 Región válida de procesado

Como se comentó en el algoritmo NTIA, los bordes de píxeles y líneas de imagen no válidas pueden influir en la medida de calidad. Estos bordes pueden deberse a sistemas de compresión que reduzcan el área de la imagen con el fin de almacenar bits de transmisión o a sistemas de procesado que deterioren algunas líneas o píxeles. Por ello, es necesario descartar un pequeño área de las secuencias, aumentando la robustez de la medida.

Es importante recordar que el descarte de una pequeña cantidad de vídeo tendrá poco impacto en la estimación de la calidad; sin embargo, la inclusión de vídeo corrupto en los cálculos de calidad puede tener un impacto mucho mayor en la estimación.

Basándose en los criterios de sencillez computacional, la región válida de procesado no se calcula dinámicamente, sino que se toman unos valores extraídos de forma empírica que aseguran unos resultados satisfactorios en la mayoría de los casos. Concretamente, la solución final consiste en eliminar 6 píxeles a izquierda y derecha, 6 líneas en la parte superior, y 4 líneas en la parte inferior para secuencias con una anchura superior a 400 píxeles. Si el tamaño es menor (por ejemplo, CIF) la región no válida se calcula reduciendo dichos valores a la mitad.

5.2.2 Ponderación por luminancia

Al estudiar los movimientos oculares y la atención visual, se analizó la particular forma del ojo para ver imágenes estáticas dando saltos de un punto de la imagen a otro mediante los denominados movimientos sacádicos. Cada movimiento es precedido por un cambio de atención a esa posición concreta. Se comprobó, por tanto, que era poco probable que en la visualización de una imagen, se hiciera un salto al vacío, es decir, que en un movimiento se fijara un punto correspondiente a un fondo o un punto que careciera de cambios abruptos de luminancia.

Basándose en esta propiedad del sistema visual humano, se desarrolla el módulo de ponderación por luminancia. En efecto, las zonas muy oscuras o muy brillantes

normalmente no atraen la atención del observador, por lo que su importancia en la medida final ha de ser menor. Para ello, los bloques correspondientes se ponderan con pesos menores.

Se usa el valor medio de la componente de luminancia como estimador de la luminancia local. Los pesos se ajustan de forma absoluta (no se tiene en cuenta la luminancia media del fotograma) al tratarse de bloques lo suficientemente grandes. La ponderación realizada es la siguiente.

En primer lugar se calcula el valor medio de la componente de luminancia local. Siendo $m(x,y)$ el píxel x e y del macrobloque:

$$\mu_x = \frac{\sum_{x=1}^{15}\sum_{y=1}^{15} m(x,y)}{16 \cdot 16}$$

Posteriormente, se ponderan los coeficientes. Para el caso de baja luminancia:

$$w_r' = \begin{cases} 0 & \mu_x \le 30 \\ w_r \cdot \left(\dfrac{\mu_x - 30}{20}\right) & 30 < \mu_x \le 50 \\ w_r & \mu_x > 50 \end{cases}$$

Siendo w_r el coeficiente de ponderación del bloque r del fotograma en cuestión.
En el caso de alta luminancia, la ponderación se realiza de la siguiente manera:

$$w_r' = \begin{cases} 0 & \mu_x \ge 240 \\ w_r \cdot \left(\dfrac{240 - \mu_x}{20}\right) & 220 \le \mu_x < 240 \\ w_r & \mu_x < 220 \end{cases}$$

En la siguiente figura, se observa la ponderación de luminancia de forma gráfica.

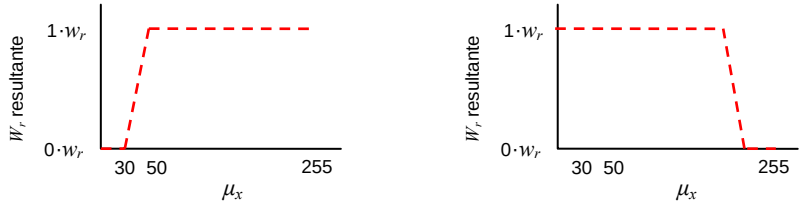


Fig. 5.8 Funciones de ponderación por luminancia

5.2.3 Actividad de bloque

El módulo de cálculo de actividad de bloque hace uso de la técnica descrita en el apartado 5.1.2, desarrollada por A.Bhat, I. Richardson y S. Kannangara en la *Robert Gordon University* de Aberdeen, Reino Unido [59] para el cálculo del índice de predicción de MOS.

Sin embargo, sus resultados no serán usados únicamente para la obtención del MOSp, sino que también se emplearán para la ponderación local del índice VSSIM. Para ello, se añaden dos parámetros adicionales, K_1 y K_2, que se obtienen a partir de la actividad espacial y temporal, respectivamente:

$$K_1 = STI / 255;$$
$$K_2 = TI / 255;$$

Siendo *STI* la Información Espacial de Textura, y *TI* la Información Temporal.

Posteriormente, se calcula la actividad del bloque como:

$$actividad\_bloque = max(K_1,K_2);$$

Y, con ella, se ponderan los coeficientes:

$$w_r' = w_r \cdot (1 - actividad\_bloque);$$

De tal forma que a medida que aumenta la actividad del bloque (ya sea espacial o temporal), el valor del coeficiente de ponderación disminuye. El impacto que dicho bloque tendrá en la evaluación perceptual final será menor, pues sufrirá un efecto de enmascaramiento.

En los bloques del primer fotograma, donde aún no se dispone de información temporal (no hay fotograma anterior para el cálculo de los bloques diferencia), la medida se realiza considerando únicamente la información espacial de textura.

En este módulo se ajustan también los pesos finales de ponderación entre MOSp y VSSIMm, de tal forma que se disminuye en un 0,01% la importancia del MOSp a favor del VSSIMm al detectar presencia de actividad (espacial o temporal) en los bloques.

5.2.4 Evaluación del movimiento

Este módulo simula los efectos de reducción visual que se producen al tener lugar cambios temporales de luminancia en la secuencia de vídeo. El mismo tipo de distorsión será más visible en secuencias estáticas que en aquellas en las que el movimiento sea elevado. Además, como ya se ha indicado, en cambios de escena, la resolución espacial puede ser reducida drásticamente sin que se llegue a percibir siempre que se restablezca la resolución original en un periodo breve de tiempo (en torno a 100 ms).

Por todo ello se implementa el bloque de evaluación de movimiento que actúa modificando los coeficientes de ponderación VSSIM a nivel de fotograma. En concreto, se otorga un mayor peso a fotogramas con bajo nivel de movimiento (donde la distorsión es más perceptible) y un menor peso a fotogramas con presencia de movimiento.

Para el cálculo, se obtiene la diferencia entre fotogramas:

$$diff_k(i,j) = \sqrt{(Y\_ref_k(i,j) - Y\_ref_{k-1}(i,j))^2}$$

siendo Y_ref_k el plano de luminancia del fotograma de referencia *k-ésimo*; e Y_ref_{k-1} el plano de luminancia del fotograma anterior. Se comienza con el fotograma 2. En el primer fotograma la ponderación es 1, al tratarse de la primera imagen que recibe el observador, que tiene gran impacto en la calidad percibida.

Posteriormente, se calcula un valor global de la diferencia para cada uno de los fotogramas de la secuencia:

$$v\_diff_k = \frac{\sum_{i=1}^{W}\sum_{j=1}^{H}|diff_k(i,j)|}{W \cdot H}$$

siendo W y H la anchura y la altura de fotograma, respectivamente. Por último se calcula el factor de ponderación para el fotograma k de la siguiente forma:

$$W_k = \left(1 - \left(\frac{v\_diff_k}{255}\right)^2\right)$$

Adicionalmente se ajustan los pesos finales de ponderación entre MOSp y VSSIMm, disminuyendo en un 0,1% la importancia del MOSp a favor del VSSIMm al superar el valor global de movimiento el umbral empírico de 15.

5.2.5 Medida de parpadeo

Este módulo evalúa la presencia del efecto de parpadeo (*flicker*) relativo a calidad de vídeo en la secuencia distorsionada. En este contexto, el flicker consiste en una variación oscilante de la calidad percibida. Al representar la progresión temporal del índice de calidad, se comprueba dicho efecto:

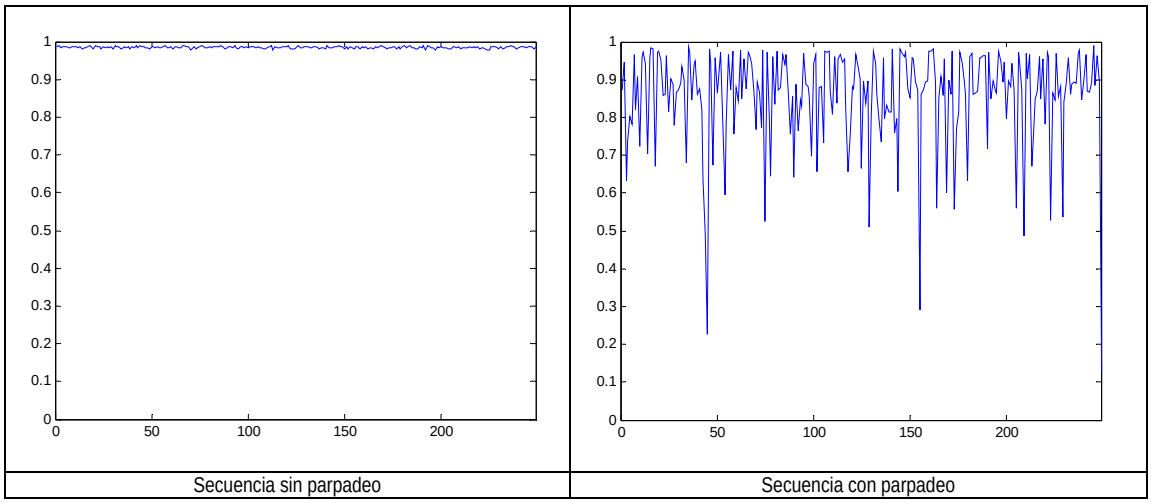


Fig. 5.9 Progresión temporal del índice de calidad en dos secuencias distintas

La presencia de parpadeo hace disminuir la valoración general de calidad de la secuencia. Sin embargo, tras las pruebas experimentales realizadas, se ha comprobado que su impacto no es tan alto como en un principio cabría imaginar. Tomando como base los resultados de dicha prueba se ajustó el módulo, que modifica tanto los coeficientes de ponderación como los pesos finales para adecuar el índice final al posible parpadeo presente en el vídeo.

En primer lugar se calcula la autocorrelación de la secuencia distorsionada. Con ello se conseguirá detectar los casos con una presencia de flicker muy alta. Se emplea el

percentil 95 de la autocorrelación. Si es mayor que 0,40 se calcula la desviación típica de la señal Q_k y, en función de ella, se ponderan coeficientes de la siguiente manera:

$$W_k' = W_k \cdot \left| \left(1 - \frac{STD}{0{,}15} \cdot Q_k \right)^2 \right|$$

donde STD es la desviación típica de Q_k. También se ajustan pesos según:

```
si KP1>0.1  y  KP2<0.9
    KP1 = KP1-0.1;
    KP2 = KP2+0.1;
```

siendo *KP1* el peso de la medida MOSp y *KP2* el peso de la medida VSSIMm. Para el resto de casos, se emplea el rango intercuartílico (percentil 75 menos percentil 25) de la secuencia temporal Q_k. Se elige el rango intercuartílico en lugar de otro estimador (como la varianza) por robustez frente a *outliers* o muestras fuera de rango.

Sin embargo surge un problema en secuencias sin parpadeo pero con un cambio abrupto en la calidad (por ejemplo, por un cambio de escena), ya que también dará un rango intercuartílico alto. Por ejemplo, en la secuencia de la figura 5.10 hay un cambio de escena aproximadamente a la mitad.

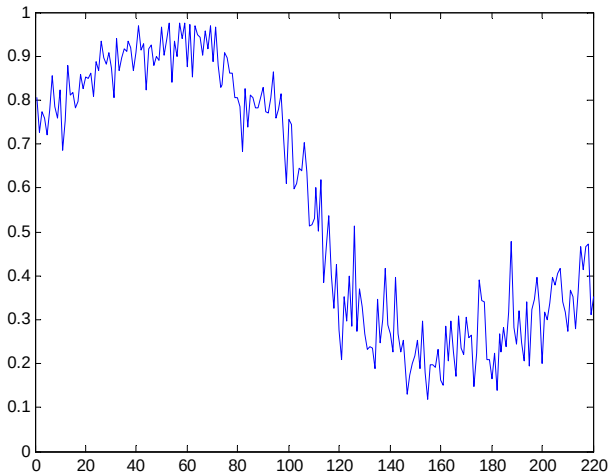


Fig. 5.10 Progresión temporal de Q_k de una secuencia con cambio de escena

Para solucionarlo, se calcula la transformada de Fourier de la secuencia, eliminando todos los coeficientes excepto los tres de más baja frecuencia. Al reconstruir, queda una descripción general de la evolución de la calidad mostrada en la figura 5.11.

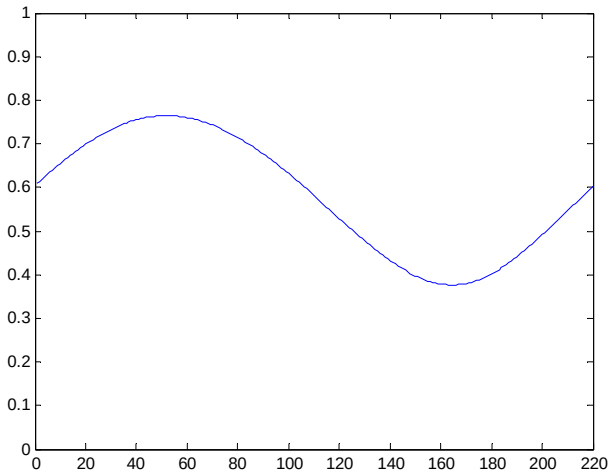


Fig. 5.11 Reconstrucción de la progresión temporal de la secuencia representada en la figura 5.10

Mientras que en las secuencias con parpadeo, la progresión de Q_k será aproximadamente plana, como se muestra en las figuras siguientes.

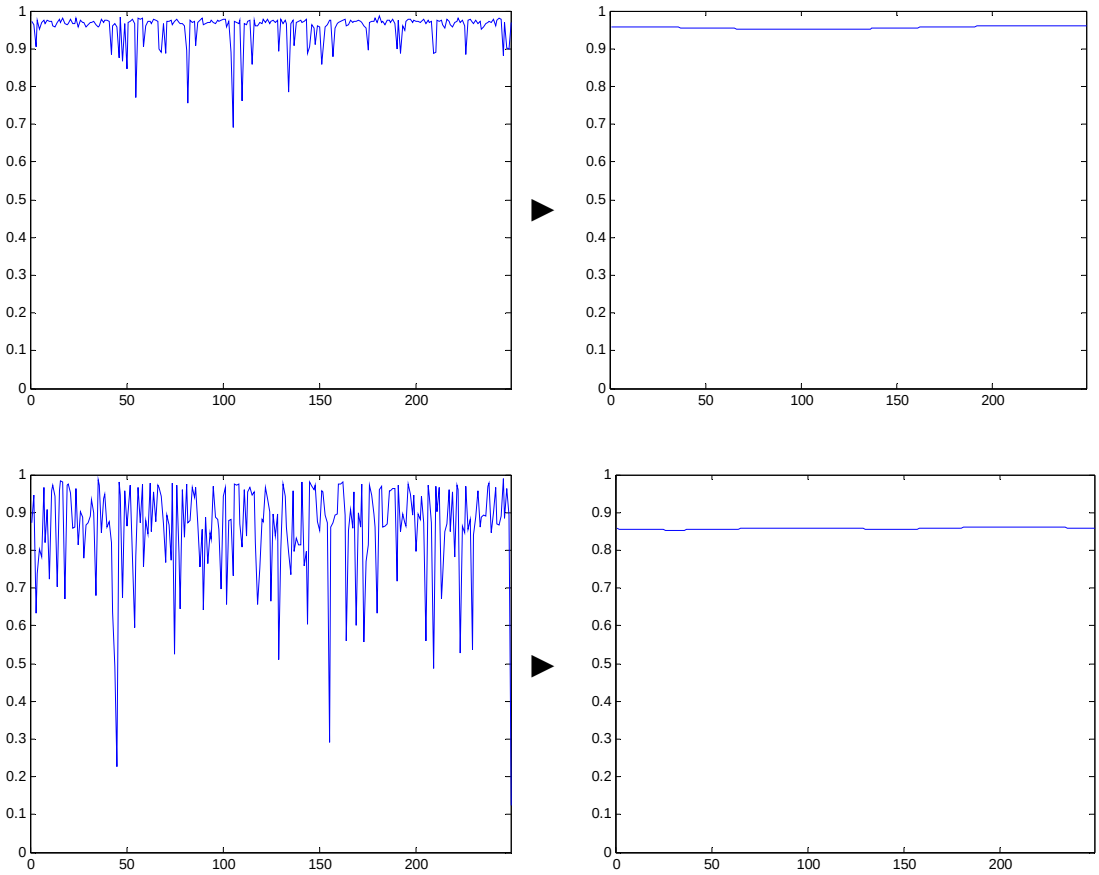


Fig. 5.12 Reconstrucciones (derecha) de las progresiones temporales mostradas a la izquierda

Por lo tanto, calculando el rango total entre valor máximo y valor mínimo de la inversa de la transformada, podemos distinguir las secuencias en las que la calidad global cambia mucho rápidamente (presencia de parpadeo) o cambia lentamente (no hay parpadeo, probablemente cambio de escena).

Tras ello, se calcula el rango intercuartílico de las secuencias y en función de este rango, se ponderan los coeficientes.

$$W_k' = W_k \cdot \left| \left(1 - \sqrt{\frac{RI}{0{,}2}} \cdot Q_k \right)^2 \right|$$

donde RI es el rango intercuartílico de Q_k.

5.3 Descripción de la aplicación

A partir del algoritmo propuesto, se ha desarrollado una aplicación que, de forma gráfica e intuitiva, permite obtener el Índice de Calidad de Vídeo explicado anteriormente. El lenguaje utilizado para desarrollar la aplicación ha sido MATLAB. La elección de este lenguaje se basa en las numerosas funciones predefinidas que incorpora MATLAB y en su sencillez para realizar pruebas al instante mediante la programación en línea de comandos, sin necesidad de compilador. Y no sólo eso, sino que la ayuda del programa y los mensajes de error proporcionan gran cantidad de información a la hora de corregir fallos y depurar el código realizado. Además para el desarrollo de la interfaz gráfica se ha usado el entorno de programación visual que ofrece MATLAB (GUIDE).

A continuación se detalla la aplicación, su funcionamiento y los resultados que produce. Para una descripción más precisa de la forma de uso, se puede consultar el manual de usuario en el anexo B.

De forma general, la estructura de entrada salida se muestra en la figura inferior.

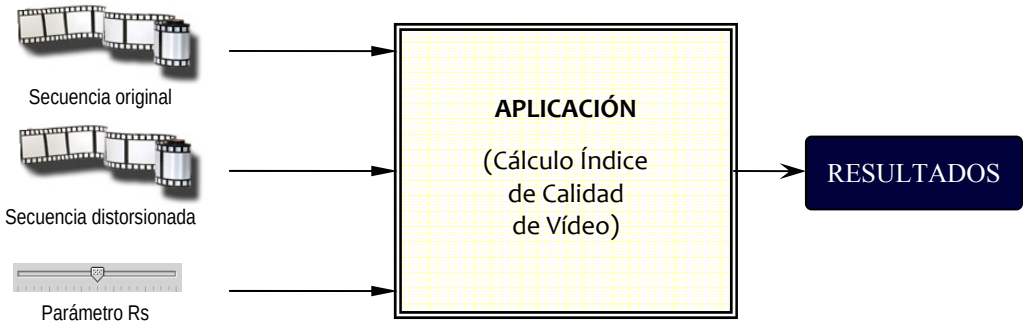


Fig. 5.13 Estructura de entrada-salida de la aplicación

La entrada se realiza mediante la primera pantalla (figura 5.14) que muestra la aplicación al ejecutar. Se han de introducir las rutas absolutas de las secuencias original y distorsionada y seleccionar el parámetro *Rs* mediante el cursor deslizante. El parámetro *Rs* (que, como se explicó, representa la densidad de muestreo) varía de 0,1 hasta 1, pudiendo el usuario seleccionar su valor en función del compromiso rapidez de ejecución – precisión en el resultado. El valor por defecto se fija a 0,1 con el que se obtienen unos resultados razonables en el menor tiempo de ejecución.

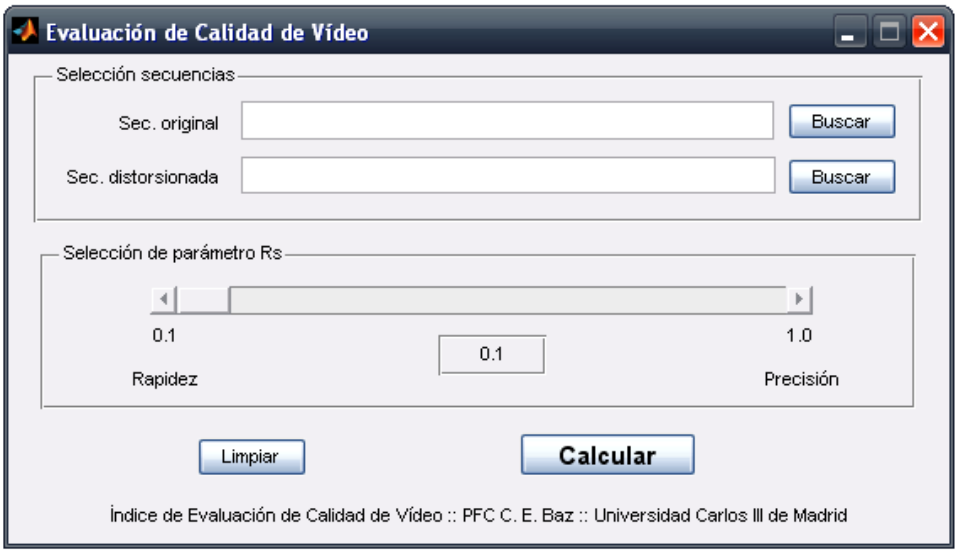


Fig. 5.14 Pantalla inicial de entrada de parámetros

En cuanto a las secuencias, el formato de entrada ha de ser .avi sin compresión. La ruta hasta ellas se puede introducir de forma manual o mediante la pantalla de búsqueda que ofrece la aplicación al pulsar sobre el correspondiente botón '*Buscar*'.

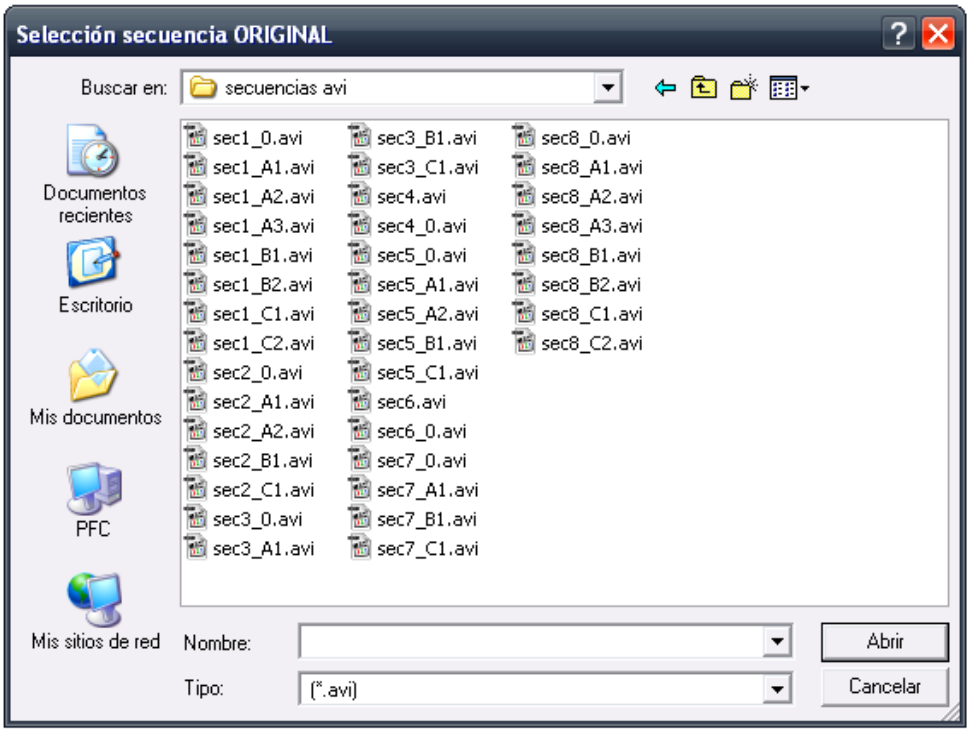


Fig. 5.15 Cuadro de búsqueda de las secuencias

Una vez se han seleccionado los parámetros de entrada se presiona el botón *Calcular* y comienza el proceso. Si alguno de los campos de secuencia está vacío se mostraría el aviso de la figura 5.16.

Fig. 5.16 Aviso de campo de ruta distorsionada vacío

Si todo es correcto, comienza el cálculo, cargando previamente las secuencias (figura 5.17). Durante la carga se pueden producir varios tipos de errores, mostrados en la figura 5.18, tras los cuales se detendría la ejecución del programa.

Fig. 5.17 Proceso de carga en memoria de secuencias

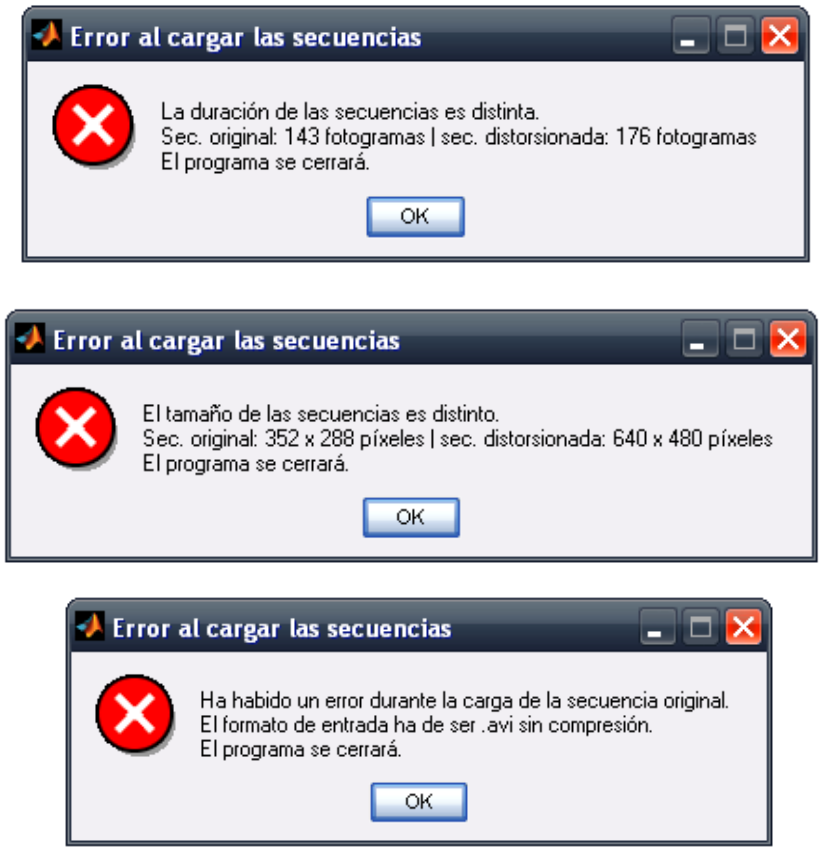


Fig. 5.18 Mensajes de error posibles durante la carga de secuencias

Si todos los campos están completos y la carga de las secuencias se realiza de forma correcta, comienza el cálculo del índice. Durante el procesado se muestra la pantalla informativa de la figura 5.19, que indica el porcentaje completado.

Fig. 5.19 Pantalla de progreso del cálculo

Finalmente, se muestra la pantalla de resultados (figura inferior).

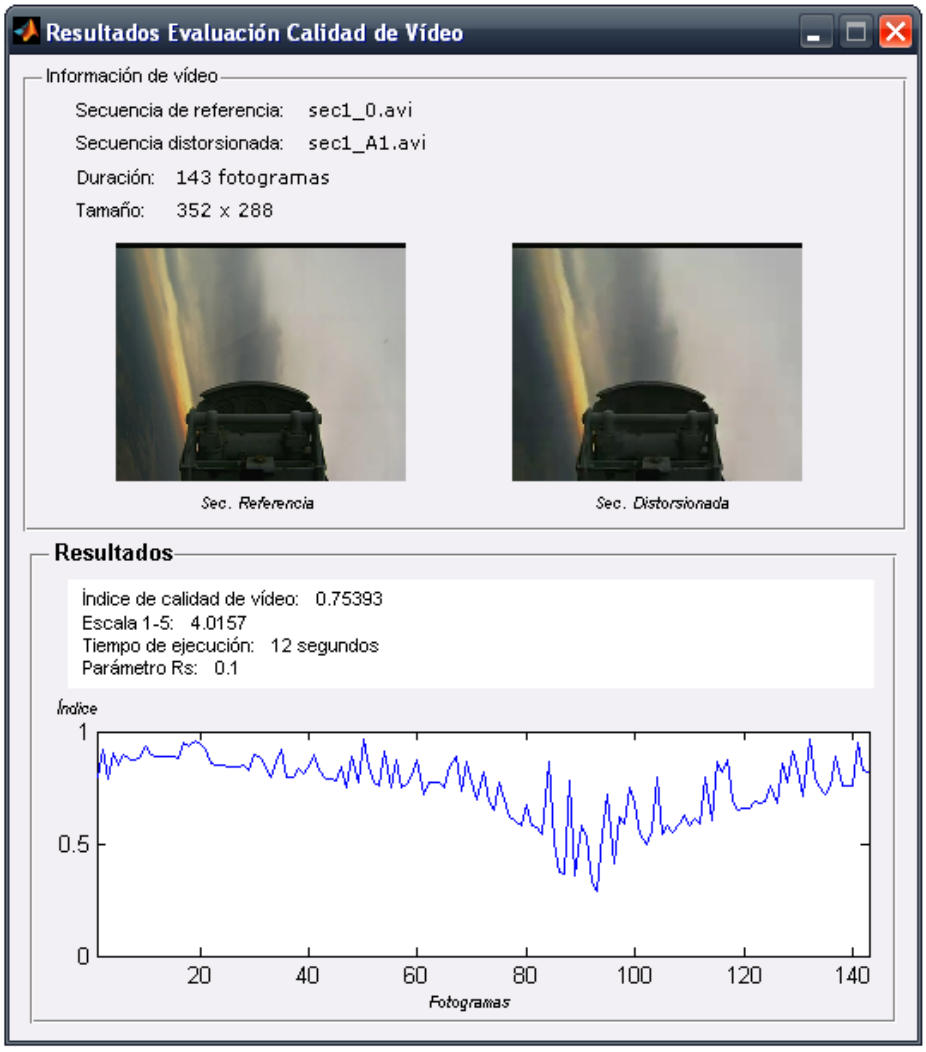


Fig. 5.20 Pantalla de resultados

La pantalla de resultados ofrece la siguiente información:

- Información de las secuencias (nombre, duración, tamaño)
- Fotogramas de muestra de ambas secuencias
- Índice de Calidad de Vídeo
- Índice en escala 1-5
- Tiempo de ejecución
- Valor del parámetro *Rs* empleado

Y, con dicha pantalla, concluye la ejecución del algoritmo. La aplicación finaliza al cerrar la ventana de resultados. Como se ha comentado anteriormente, una descripción más precisa de la ejecución y uso de la aplicación se realizará en el anexo B.

6

Prueba subjetiva y resultados

El fin de este capítulo es analizar los resultados obtenidos para determinar si se cumplen los objetivos marcados y evaluar la eficacia tanto del sistema propuesto como del resto de algoritmos analizados.

Para comprobar si los resultados de los algoritmos se ajustan a la opinión subjetiva, fue necesario realizar una prueba experimental con observadores humanos en la que se obtuvo la calidad de un conjunto de secuencias. Estas secuencias no sólo se han empleado para ajustar los parámetros de la medida propuesta, sino que también han servido para la posterior evaluación de algoritmos.

6.1 Desarrollo de la prueba subjetiva

En este primer apartado se realizará una descripción de la prueba subjetiva realizada durante la ejecución de este proyecto. El objetivo del experimento era la valoración subjetiva de la calidad de un número determinado de secuencias de vídeo.

En efecto, la manera más segura y fiable de determinar la calidad de las secuencias de vídeo es mediante la evaluación subjetiva, ya que los humanos serán los receptores finales.

Con esta prueba se obtuvo una puntuación media de opinión (MOS) que caracteriza dicha valoración subjetiva de la calidad de cada una de las secuencias mostradas.

La realización del experimento se basó principalmente en la recomendación P.910 de la ITU-T [60], aunque también se consideraron aspectos de la recomendación BT.500-11 de la ITU-R [61].

6.1.1 Método de prueba

De entre los métodos disponibles, se eligió para el experimento el Índice por Categorías de Degradación *(Degradation Category Rating)*, o Método de escala de degradación con Doble Estímulo, *DSIS* (ITU-R BT.500-11). Proporciona una evaluación precisa de la transparencia o fidelidad del sistema usado para procesar las secuencias, pues el observador compara directamente la secuencia procesada con la referencia. Sirve además para evaluar métricas de tipo FR (Full Reference), en las que se dispone de la señal original para realizar la predicción.

En este método, las secuencias se presentan en pares: el primer estímulo de cada par es siempre la referencia y el segundo la secuencia de prueba. Los observadores han de estar avisados de esta disposición.

El patrón de tiempos a seguir es el siguiente:

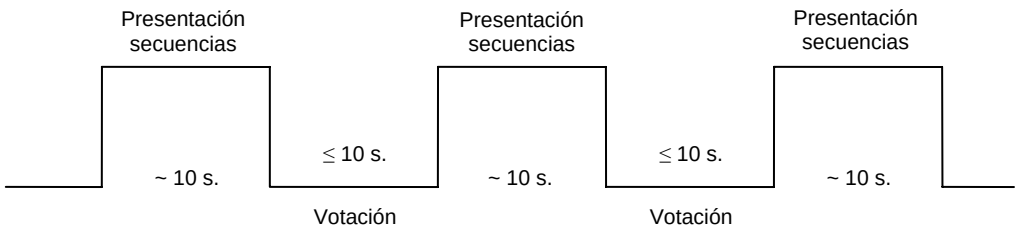


Fig. 6.1 Patrón de tiempos del método DSIS

Los sujetos evalúan la degradación del segundo estímulo en relación con la referencia con la siguiente escala de cinco niveles.

5 Imperceptible
4 Perceptible, pero no molesta
3 Ligeramente molesta
2 Molesta
1 Muy molesta

No obstante, en el experimento se usó el modo *Simultaneous Presentation* (SP) en el que se visualizan las secuencias de prueba y de referencia simultáneamente en el mismo monitor. La referencia se sitúa a la izquierda de la pantalla, mientras que la secuencia distorsionada se sitúa a la derecha. Con ello se obtienen una serie de ventajas como la reducción de la duración de la prueba, una evaluación de diferencias más sencilla para los observadores así como la mayor atención de éstos al reducir el número de pruebas a la mitad.

Para poder llevar a cabo el DCR en modo SP, se han de tener en cuenta las siguientes consideraciones:

Uso de formato reducido
Las dos secuencias han de estar perfectamente sincronizadas
Fondo gris 50%
Distancia de observación de 1H a 8H (H: altura de las imágenes)
Monitor de 14" como mínimo

6.1.2 Secuencias empleadas

Se emplearon dos codificadores para las secuencias usadas en el experimento: el software de referencia H.264 y un codificador con consideraciones perceptuales. En total se dispuso de 30 secuencias codificadas a distinta calidad que provenían de ocho secuencias fuente. Las secuencias son de tamaño CIF y 25 fotogramas por segundo.

Las secuencias fuente empleadas son las siguientes:

Fig. 6.2 Secuencias fuente empleadas en la prueba subjetiva

Para evaluar la información perceptual espacial y la información perceptual temporal de las escenas, se hace uso de las siguientes medidas (ITU-T P.910):

Medida de información espacial perceptual:

$$SI = \max_{tiempo}\left\{ std_{espacio}\left[Sobel(F_n)\right]\right\}$$

F_n Fotograma (plano de luminancia) en el instante n

Medida de información temporal perceptual:

$M_n(i, j)$ Diferencia entre los valores de los píxeles (plano de luminancia) con la misma localización espacial pero en fotogramas sucesivos.

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j)$$

$F_n(i, j)$ Píxel en la fila i-ésima y en la columna j-ésima del fotograma n

$$TI = \max_{tiempo}\left\{ std_{espacio}\left[M_n(i,j)\right]\right\}$$

(Más movimiento en fotogramas adyacentes resultará en un valor mayor de TI)

Por lo tanto, las secuencias se clasifican según:

- *TI=0:* escenas estáticas o con muy poco movimiento
- *TI alto:* escenas con mucho movimiento
- *SI=0* escenas con muy poco detalle espacial
- *SI alto:* escenas con mucho detalle espacial

Según lo cual, la información espacial (SI) y la información temporal (TI) de las secuencias empleadas es la mostrada en las siguientes tablas.

	Airshow	Bohemia	Bus	Container
SI	77.7099	100.6374	133.9806	125.9194
TI	9.2636	25.8199	32.9130	7.4717

	Football	LOTR	Star Wars	Stefan
SI	126.6030	66.6864	88.3396	127.2813
TI	29.7539	13.8572	28.8235	40.0050

Tabla 6.1 Información espacial y temporal de las secuencias empleadas

De forma gráfica, en el plano SI-TI, las secuencias se disponen de la siguiente forma:

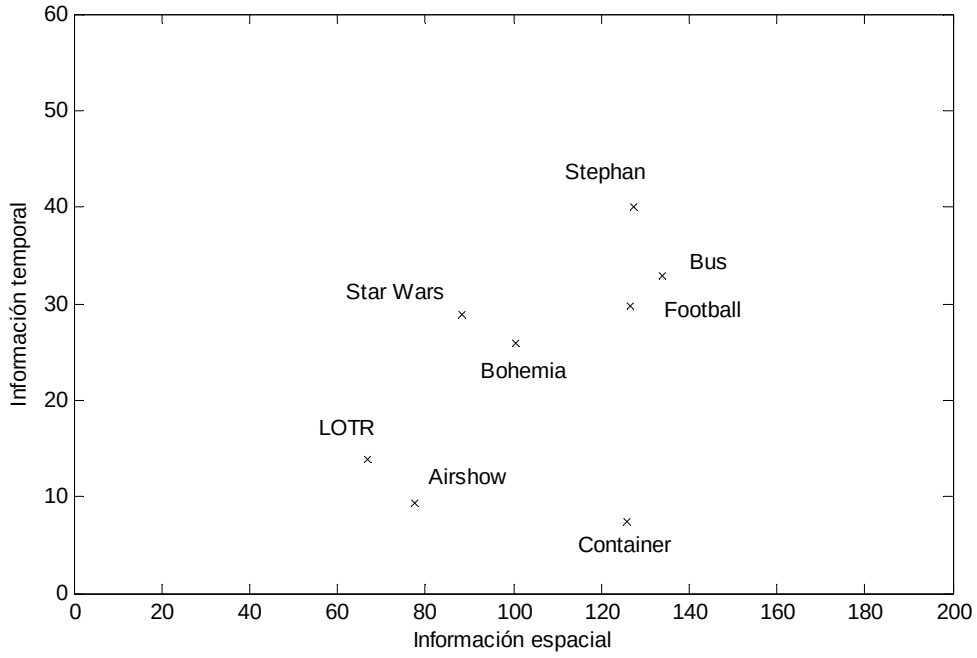


Fig. 6.3 Gráfico SI-TI secuencias fuente

6.1.3 Diseño experimental

En el diseño experimental se siguió la recomendación P.910. Para el desarrollo de la prueba se emplearon monitores TFT de 19" con una resolución de 1280 x 720 píxeles. El método DCR se llevó a cabo, como se ha mencionado, mediante modo SP, con un color de fondo de gris al 50% (Y=U=V=128, U y V sin signo).

La distancia de observación pudo variar de 1 a 8 H, siendo H la altura de la imagen. Para conseguir una iluminación de fondo menor de 20 lux; se apagaron todas las fuentes de luz y se cerró cualquier posible ventana de la sala de pruebas.

Cada sesión de pruebas se preparó para no exceder una duración de quince minutos aproximadamente. Hay que tener en cuenta que durante la realización de las pruebas se incluyeron dos reiteraciones (repeticiones de condiciones idénticas) para calcular la fiabilidad individual de cada sujeto.

Por otro lado, se incluyeron también 3 presentaciones al comienzo de cada prueba que no se tuvieron en cuenta en el análisis estadístico de los resultados y que sirven para estabilizar la opinión de los observadores.

Por lo tanto, la duración de la prueba se estimó en:

- 30 pares de secuencias
- 2 pares de secuencia de repeticiones
- 3 pares de secuencia de entrenamiento

35 · 8 s. duración + 35 · 10 s. votación = 630 s.

Con una duración media de secuencia de 8 segundos, se obtiene una duración de prueba de 10 minutos y 30 segundos. A ello hay que sumar el tiempo empleado en las pruebas visuales (detalladas más adelante).

6.1.4 Observadores

Según la recomendación, el número de observadores ha de ser de 4 a 40 (difícilmente se obtienen mayores ventajas con más de 40 sujetos), con un mínimo recomendado de 15

observadores. Estas personas no deben intervenir directamente en evaluaciones de calidad de imagen como parte de su trabajo habitual y no han de ser evaluadores experimentados.

Para la prueba se contó con un total de 47 personas no expertas. Se tomaron más sujetos del máximo recomendado por posibles rechazos de observadores (por agudeza visual o por votación no válida). En efecto, antes de comenzar la sesión de pruebas, se comprobó la visión de los observadores. Con respecto a la agudeza visual, no debían cometerse errores en la línea 20/30 del diagrama de ojo normalizado de Snellen (figura 6.4.a). Por lo que se refiere al color, no se debían perder más de 2 placas de Ishihara de un total de 12 (figura 6.4.b).

De los 47 observadores se descartaron 5 por los siguientes motivos:

- Test Isihara no apto (2 sujetos)
- Repeticiones no concordantes
- Distribución inadecuada de la votación
- Test Snellen no apto

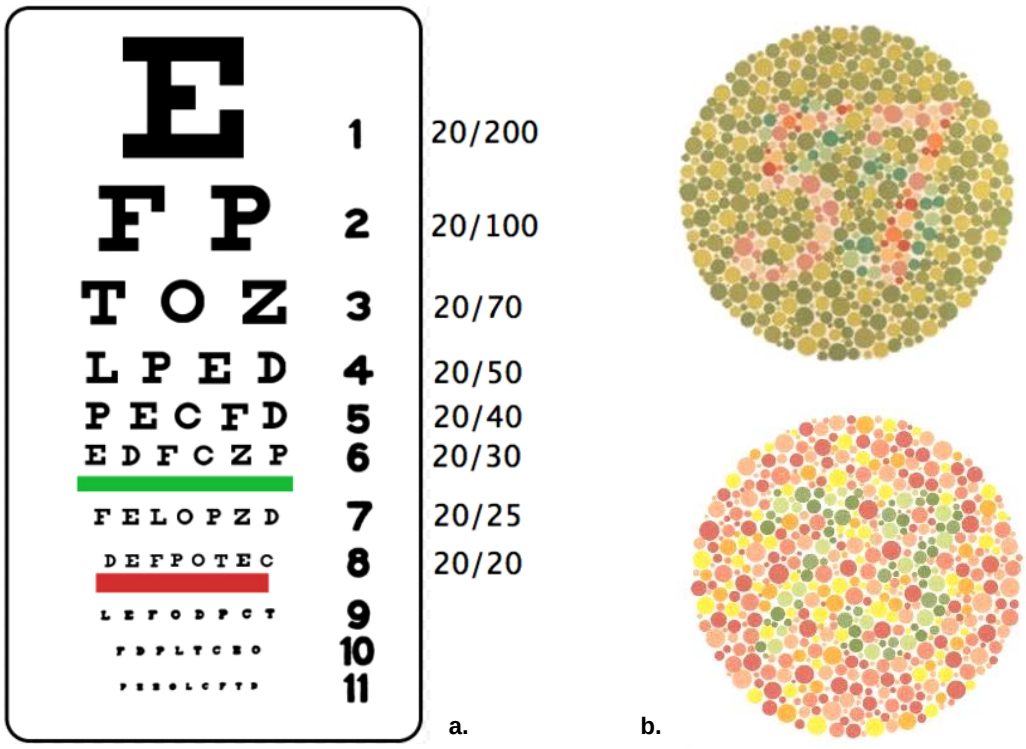


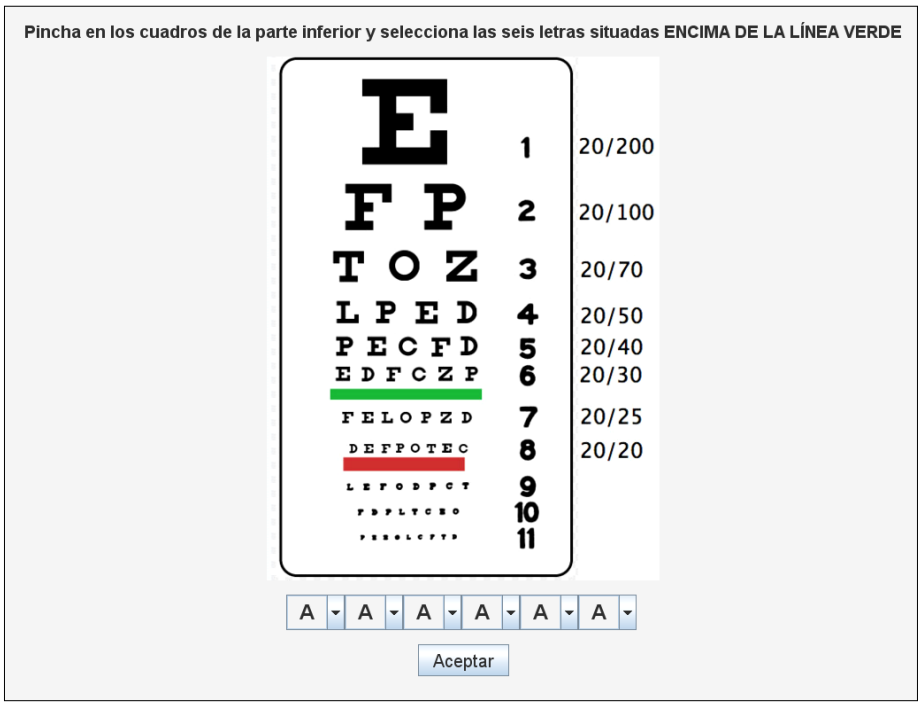
Fig. 6.4 Pruebas visuales: **a.** Test Snellen **b.** Dos placas Isihara de ejemplo

6.1.5 Software de control de la prueba

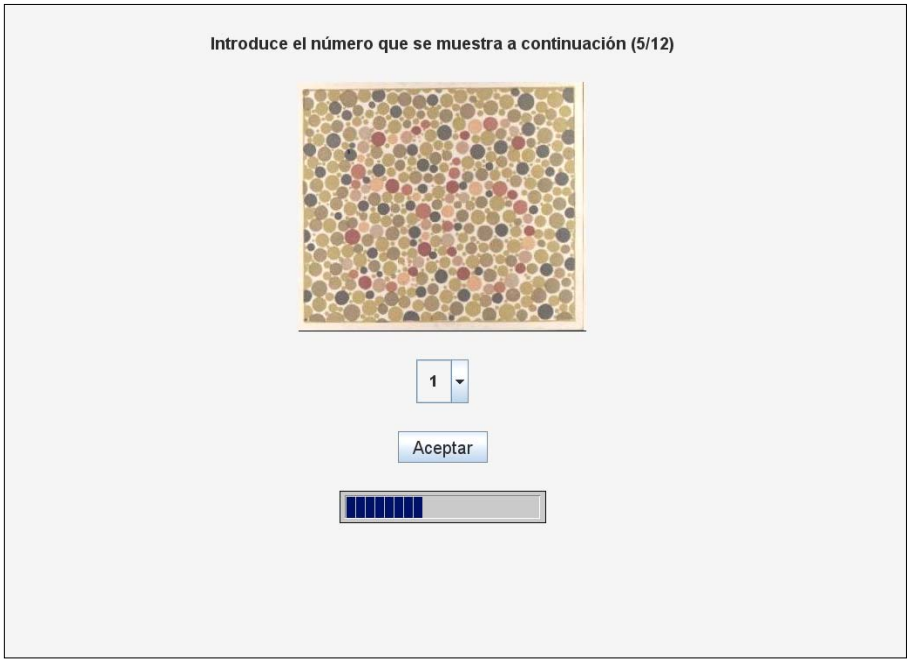
Para almacenar y reproducir el contenido vídeo se utilizaron PCs, así como un software de propósito específico. Dicho software se desarrolló en Java (versión 1.6) haciendo uso de la JMF 2.1.1e. Proporciona una interfaz en la que los observadores evaluaron cada una de las secuencias mostradas. Permite además valorar la agudeza visual de los sujetos mediante los tests de Snellen y de Ishihara.

Para sincronizar las secuencias en el modo DCR-SP, antes de comenzar la reproducción se determinan y se adquieren los recursos necesarios para ambos reproductores java, llevándolos al estado *prefetched*. Una vez ambos reproductores están listos para la reproducción, éstos se inician de forma simultánea.

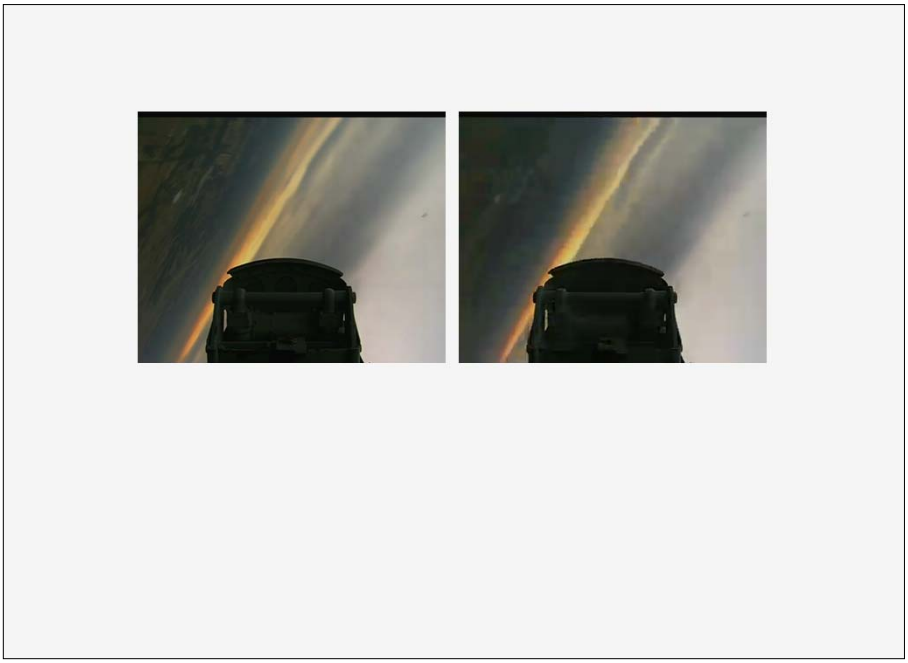
A continuación se muestran unas capturas de distintas pantallas del software:



a.



b.



c.

Fig. 6.5 Pantallas de muestra del software de control: **a.** Test Snellen **b.** Placa de test Isihara **c.** Prueba DCR en modo SP

La votación se realiza mediante una pantalla de selección mostrada a los observadores a la finalización de cada par de secuencias.

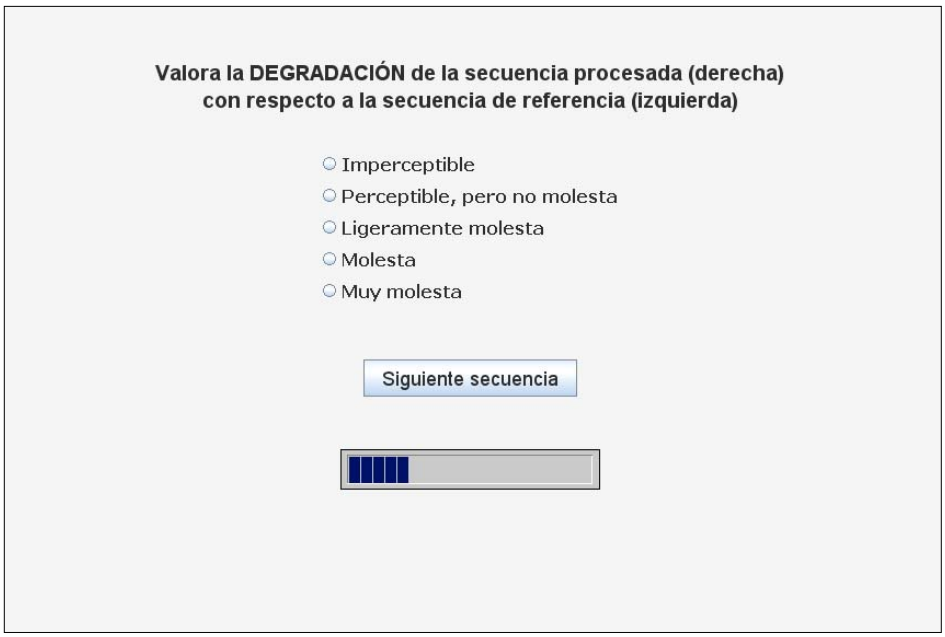


Fig. 6.6 Pantalla de votación del software de control

6.2 Resultados obtenidos y comparación de algoritmos

Una vez obtenido el conjunto de votaciones de los 42 observadores válidos de la prueba, se calculan los valores medios. Para ello se hace uso de la puntuación media de opinión, que se obtiene de la siguiente forma:

$$MOS_k = \frac{1}{N} \cdot \sum_{i=1}^{N} \sum_{<k>} u_{ikr}$$

donde u_{ikr} es la nota del observador i para la secuencia k en la repetición r, y N es el número de observadores (47).

Adicionalmente, cada una de las notas medias tendrá un intervalo de confianza asociado que se obtiene a partir de la desviación típica y el tamaño de cada muestra.

Según la recomendación ITU-R BT.500-11, se utiliza un intervalo de confianza del 95%, que viene dado por.

$$[MOS - \delta_{kr}, MOS + \delta_{kr}]$$

donde:

$$\delta_{jr} = 1{,}96 \cdot \frac{S_{kr}}{\sqrt{N}}$$

La desviación típica de cada presentación, S_{kr} viene dada por:

$$S_{kr} = \sqrt{\sum_{i=1}^{N} \frac{(MOS - u_{ikr})^2}{(N-1)}}$$

Estos cálculos asumen una distribución normal de la población. En efecto, si representamos el histograma de las votaciones de una secuencia cualquiera de las evaluadas en la prueba, resulta:

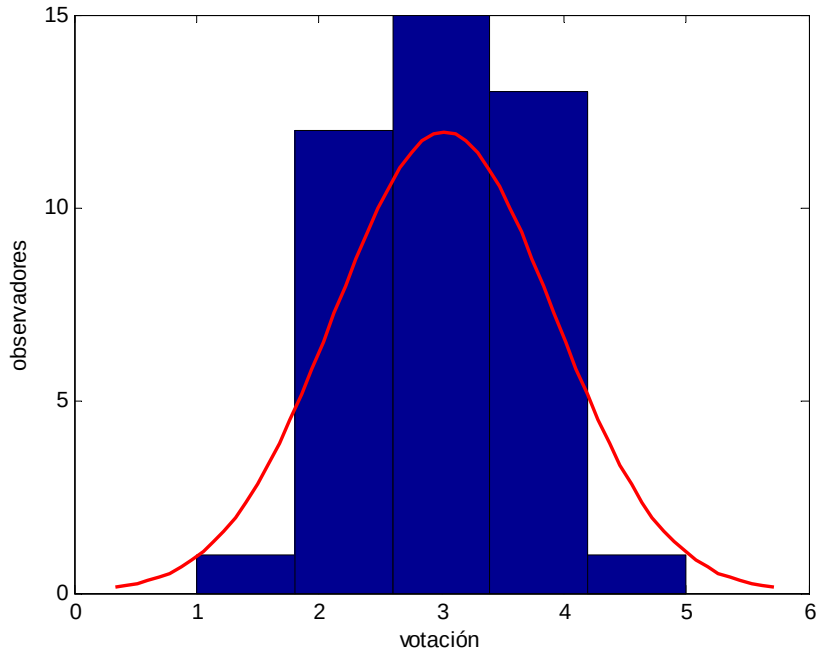


Fig. 6.7 Distribución de votación de la secuencia *sec1_C1*. En rojo, el ajuste a la distribución normal

Por lo tanto, se puede afirmar que con una probabilidad del 95%, el valor absoluto de la diferencia entre la nota media experimental y la nota media "verdadera" (para un número de observadores muy elevado) es menor que el intervalo de confianza del 95%.

Según lo anterior, los resultados obtenidos tras el procesamiento de los datos de la prueba subjetiva se resumen en la siguiente tabla:

Secuencia	MOS	$\delta_{95\%}$	Secuencia	MOS	$\delta_{95\%}$
sec1_A1	3,0476	0,30209	sec5_A1	3,3571	0,27067
sec1_A2	4,9523	0,06441	sec5_A2	4,0952	0,19588
sec1_A3	4,7857	0,14055	sec5_B1	3,7976	0,18973
sec1_B1	2,5476	0,28156	sec5_C1	3,5476	0,25732
sec1_B2	4,7142	0,13662	sec6	4,4761	0,18942
sec1_C1	3,0238	0,26798	sec7_A1	2,9285	0,25037
sec1_C2	4,7857	0,10509	sec7_B1	2,5714	0,26516
sec2_A1	1,7619	0,21698	sec7_C1	2,8095	0,24902
sec2_A2	2,1667	0,27144	sec8_A1	2,8809	0,25732
sec2_B1	2,2142	0,24280	sec8_A2	4,3095	0,21373
sec2_C1	2,1667	0,27144	sec8_A3	4,8809	0,09794
sec3_A1	3,1904	0,24012	sec8_B1	3,6190	0,17401
sec3_B1	3,0714	0,27523	sec8_B2	4,8095	0,11875
sec3_C1	3,0119	0,19932	sec8_C1	3,4523	0,23979
sec4	4,0952	0,27058	sec8_C2	4,4047	0,21947

Tabla 6.2 Resultados prueba subjetiva

Por otro lado, se obtienen las estimaciones tanto de los algoritmos analizados como de la aplicación propuesta. En el caso de la medida NTIA se hace uso del software de referencia proporcionado por los autores. El cálculo de los algoritmos de Watson modificado y VSSIM se realiza mediante el Software *MSU Video Quality Measurement Tool* [62] del grupo de vídeo *MSU Graphics & Media Lab* de la Universidad Estatal de Moscú [63].

A continuación se detallan los valores obtenidos para cada una de las secuencias. Se diferencian dos grupos: secuencias de entrenamiento y secuencias de evaluación empleadas en el diseño de la aplicación. Los grupos se eligen con una actividad espacio temporal distribuida y con una proporción de 30% secuencias de entrenamiento y 70% de evaluación.

Secuencias de entrenamiento	MOS	Watson mod.	VSSIM	NTIA	Medida propuesta
sec4	4,0952	1,1064	4,7123	4,2380	4,4256
sec5_A1	3,3571	1,0358	4,0573	3,1596	3,2742
sec5_A2	4,0952	1,6072	4,1747	3,3664	3,6311
sec5_B1	3,7976	1,0469	4,0905	3,2648	3,3069
sec5_C1	3,5476	1,1013	4,0926	3,2512	2,9639
sec6	4,4761	1,0862	4,6772	4,1944	4,0959
sec7_A1	2,9285	1,0129	4,3198	2,2496	1,4495
sec7_B1	2,5714	1,1428	4,3308	2,2628	1,5212
sec7_C1	2,8095	1,1395	4,3292	2,2668	1,1507

Secuencias de evaluación	MOS	Watson mod.	VSSIM	NTIA	Medida propuesta
sec1_A1	3,0476	1,3723	4,8419	3,6256	4,0509
sec1_A2	4,9523	3,6938	4,9657	4,8564	4,9064
sec1_A3	4,7857	3,8024	4,9703	4,8439	4,9180
sec1_B1	2,5476	1,2961	4,8346	3,5156	4,0361
sec1_B2	4,7142	3,6733	4,9638	4,8308	4,8921
sec1_C1	3,0238	1,0029	4,8384	3,5408	4,0407
sec1_C2	4,7857	3,6052	4,9592	4,8272	4,8856
sec2_A1	1,7619	1,0682	4,1864	2,3600	1,9037
sec2_A2	2,1667	1,0464	4,3452	2,6756	2,4790
sec2_B1	2,2142	1,0517	4,1684	2,2036	1,5397
sec2_C1	2,1667	1,0285	4,2048	2,3928	1,9399
sec3_A1	3,1904	1,0183	3,8646	2,9264	3,0836
sec3_B1	3,0714	1,6919	3,8048	2,7952	3,0249
sec3_C1	3,0119	1,9844	3,8037	2,6720	2,7414
sec8_A1	2,8809	1,1508	4,3915	3,6924	4,4397
sec8_A2	4,3095	1,3462	4,8348	4,6716	4,7116
sec8_A3	4,8809	1,4208	4,8713	4,7512	4,7696
sec8_B1	3,6190	1,1365	4,3791	3,7472	3,4827
sec8_B2	4,8095	1,5631	4,8398	4,6936	4,7135
sec8_C1	3,4523	1,6615	4,2674	3,4612	3,2186
sec8_C2	4,4047	1,8431	4,8141	4,6680	4,6531

Tabla 6.3 Estimaciones de los diferentes algoritmos para secuencias de entrenamiento y de evaluación

Comparando las valoraciones subjetivas de las secuencias y las estimaciones de los algoritmos, es posible evaluar la eficacia de cada uno de los métodos analizados así como de la aplicación propuesta. Para ello se hará uso del coeficiente de correlación de Pearson, que es un índice estadístico que mide la relación lineal entre dos variables. El índice

discurre en valor absoluto entre 0 y 1, indicando la magnitud de la relación. Los valores menores que cero indican relación lineal negativa. En este contexto, un coeficiente igual a 1 indicaría una estimación de calidad igual a la valoración subjetiva.

El cálculo del coeficiente se realiza de la siguiente manera:

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

donde σ_{XY} es la covarianza de X e Y, y σ_X, σ_Y son las desviaciones típicas de las distribuciones marginales.

Con todo ello, los resultados que se obtienen en la evaluación de los algoritmos se muestran en la tabla 6.4

Secuencias de entrenamiento:

	VSSIM	Watson mod.	NTIA	Medida propuesta
Correlación de Pearson	**0,3442**	**0,2764**	**0,9395**	**0,9330**

Secuencias de evaluación:

	VSSIM	Watson mod.	NTIA	Medida propuesta
Correlación de Pearson	**0,6584**	**0,6957**	**0,9395**	**0,8600**

Tabla 6.4 Comparación de los distintos algoritmos

Adicionalmente se indica la relación de muestras fueras de rango (*outliers ratio*) que es una medida de la consistencia de la predicción. Un dato se considera fuera de rango si la diferencia entre el valor estimado y el valor subjetivo excede ±2 veces la desviación típica de los resultados subjetivos. Se indica así mismo el tiempo de ejecución, calculado para una secuencia CIF de 7s. de duración en un PC con procesador de 2 x 2,20GHz, 2GB de RAM y con SO multiproceso. En el cálculo de la estimación de la medida propuesta se ha usado un valor de *Rs* de 0,1. Para comparar resultados, el tiempo del algoritmo NTIA no incluye alineación espacial ni temporal.

	VSSIM	Watson mod.	NTIA	Medida propuesta
Relación de Outliers	**0,4667**	**0,8333**	**0,0667**	**0,2333**
Tiempo de ejecución	**5 s.**	**6 s.**	**54 s.**	**8 s.**

Tabla 6.5 Parámetros adicionales de los distintos algoritmos

A continuación se representan los diagramas de dispersión para las dos medidas que obtienen mejores resultados: la NTIA y la propuesta. Se indica el ajuste lineal para cada uno de los modelos. El algoritmo perfecto sería aquel en el que todos sus puntos coincidieran con la recta (*Estimación = MOS*).

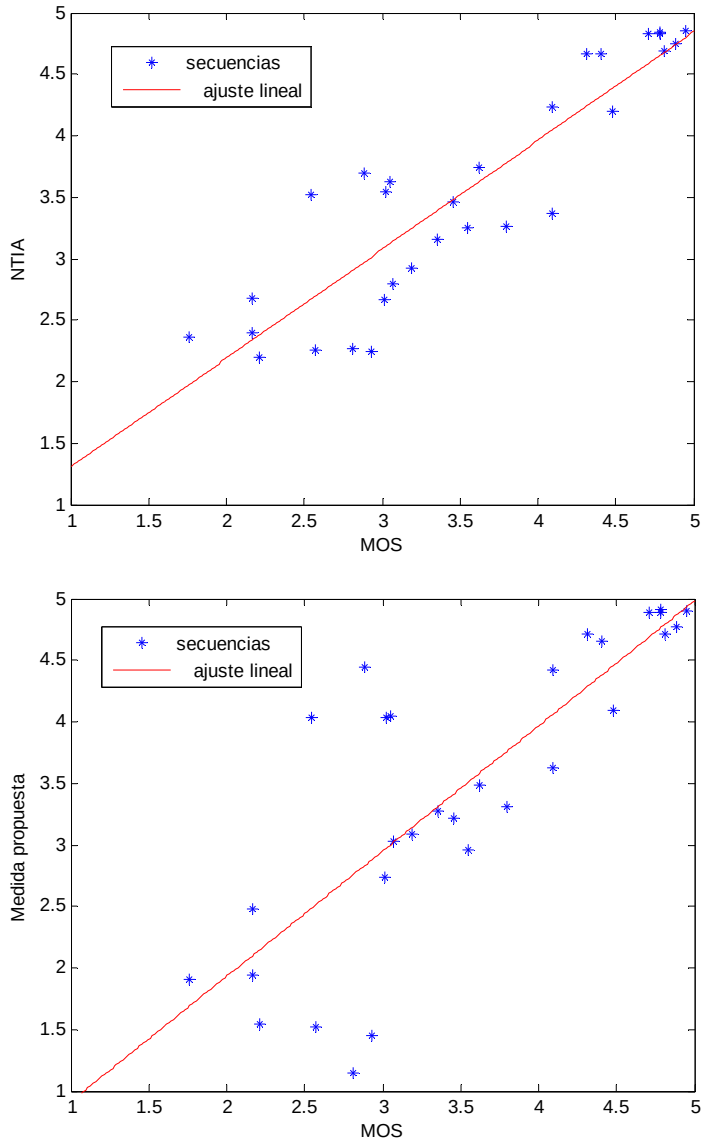


Fig. 6.8 Dispersión de los algoritmos NTIA (arriba) y propuesto (abajo)

Los algoritmos VSSIM y Watson modificado ofrecen un tiempo de ejecución realmente bajo, pero con unos resultados muy malos. El algoritmo que mejores resultados proporciona es el NTIA (correlación de 0,9395), sin embargo se puede comprobar que el tiempo de ejecución es muy elevado: 54 segundos para una secuencia de 7 segundos de duración. La medida propuesta se encuentra en un punto intermedio, elevando la correlación con respecto a los dos primeros algoritmos (0,8600), sin aumentar demasiado el tiempo de ejecución (8 segundos). En cuanto a la relación de muestras fuera de rango, en la medida propuesta se reduce (0,2333) con respecto a VSSIM sin llegar a los valores de la NTIA (0,0667).

Adicionalmente, para comprobar el rendimiento de la aplicación en otros formatos así como su robustez ante otros tipos de distorsiones, se realizó una pequeña prueba reducida de 15 personas consistente en la evaluación de 21 secuencias D1 (720 x 576 píxeles, 25 fotogramas por segundo). Las secuencias provenían de 4 secuencias fuente mostradas en la figura 6.9. En la tabla 6.6 se indica la información espacio temporal de cada una de ellas.

sec_1 sec_2 sec_3 sec_4

Fig. 6.9 Secuencias D1 empleadas en la evaluación

	sec_1	sec_2	sec_3	sec_4
SI	95.5	68.9638	99.1062	83.9251
TI	42.34	26.2755	17.8025	33.1238

Tabla 6.6 Información espacio temporal de las secuencias D1

Las 21 secuencias de prueba se generaron mediante codificación H264, suma de constante, aumento de contraste y desenfoque de 6 píxeles. Hallando las estimaciones de los algoritmos para dichas secuencias y analizando los datos, se obtienen los siguientes resultados:

	VSSIM	Watson mod.	NTIA	Medida propuesta
Correlación de Pearson	**0,7463**	**0,5413**	**0,8631**	**0,8467**

Tabla 6.7 Comparación de los distintos algoritmos para las secuencias D1

Los tiempos de ejecución (calculados con el mismo equipo que en el caso de las secuencias CIF) se muestran a continuación. Al igual que en el caso anterior, para poder comparar resultados, el tiempo del algoritmo NTIA no incluye alineación espacial ni temporal.

	VSSIM	Watson mod.	NTIA	Medida propuesta
Tiempo de ejecución	**29 s.**	**25 s.**	**348 s.**	**52 s.**

Tabla 6.8 Comparación de tiempos de ejecución de los distintos algoritmos para las secuencias D1

Se puede comprobar que la medida propuesta sigue manteniendo una correlación alta; menor que la medida NTIA, pero más próxima que en el caso anterior. El tiempo de ejecución se eleva con respecto a VSSIM o Watson modificado, pero sin llegar al valor de 348 segundos del algoritmo NTIA.

7

Conclusiones y trabajos futuros

En este capítulo se pretende dar una visión general de los objetivos alcanzados mediante la realización del presente proyecto, así como de las posibles ampliaciones que se podrían llevar a cabo tomando como base el trabajo realizado. Por tanto, en primer lugar se expondrán las conclusiones generales que se han obtenido de la implementación de la medida de calidad así como de la evaluación del resto de métricas. En segundo lugar, se van a indicar posibles recomendaciones para la optimización de la medida propuesta e investigaciones futuras.

7.1 Conclusiones

El objetivo fundamental de este proyecto ha sido el diseño e implementación de una medida objetiva de calidad perceptual de vídeo. Este propósito se ha logrado mediante un importante estudio e investigación de las medidas actualmente usadas con el fin de analizar las técnicas en las que se basan y la filosofía empleada. Adicionalmente se ha realizado una prueba subjetiva con cuyos resultados se ha ajustado la métrica propuesta y se han evaluado las distintas medidas, comparando su eficacia.

Además, la creación de un entorno gráfico (detallado en el Anexo B) proporciona una herramienta muy útil para la fácil y rápida evaluación de la calidad perceptual mediante la medida propuesta. Una vez finalizado el proyecto, a continuación se exponen las conclusiones generales que se han extraído en el desarrollo del mismo.

Analizando las técnicas estudiadas, se podía observar una importante diferencia entre ellas. Por un lado, la medida de Watson modificada proporcionaba un tiempo de ejecución muy bajo, pero con unos resultados no muy precisos. Por otro lado, la medida NTIA ofrecía unos resultados muy buenos, pero a costa de una carga computacional muy elevada que provocaba un tiempo de ejecución bastante importante.

Ante esto, los requisitos de diseño del algoritmo se centraban en lograr una medida sin excesiva carga computacional que obtuviera unos resultados razonables. Para ello se han implementado distintos tipos de algoritmos y se han empleado diversas técnicas hasta que la correlación con el MOS fue lo suficientemente buena como para considerar cumplidos los objetivos. Si se comprueban los resultados obtenidos en la evaluación de algoritmos tras la prueba subjetiva (tablas 6.4 y 6.5), se observa que los objetivos propuestos se han conseguido. A modo de recordatorio, se indican a continuación dichos resultados.

	VSSIM	Watson mod.	NTIA	Medida propuesta
Correlación de Pearson	0,6584	0,6957	0,9395	0,8600
Tiempo de ejecución	5 s.	6 s.	54 s.	8 s.

Tabla 7.1 Comparación de los distintos algoritmos (recordatorio)

En efecto, la medida propuesta se encuentra en un punto intermedio, elevando la correlación con respecto a los dos primeros algoritmos, sin aumentar en gran medida el tiempo de ejecución. No obstante, es necesario recordar que las medidas mostradas se obtuvieron con un valor de *Rs* de 0,1. Como se indicó en la descripción del sistema, el parámetro *Rs* representa la densidad de muestreo por fotograma, pudiendo ser modificado por el usuario desde 0,1 hasta 1 en función del compromiso entre rapidez de ejecución (Rs bajo) y precisión en los resultados (*Rs* alto).

Adicionalmente, la prueba subjetiva y su posterior comparación con el resto de algoritmos analizados, ha proporcionado una importante información acerca del estado actual de las medidas objetivas de calidad perceptual de vídeo. Además se ha generado una base de datos con evaluaciones subjetivas de un conjunto determinado de secuencias de vídeo que puede ser utilizada en un futuro para posteriores investigaciones y análisis de futuras técnicas y algoritmos de evaluación de calidad.

7.2 Líneas de trabajo futuras

En la elaboración de un sistema de medida de calidad perceptual es fundamental el modelado que se haga del SVH. Por lo tanto, las posibles mejoras y campos de investigación futura se basan, precisamente, en la implementación de nuevos mecanismos que formen parte del proceso de percepción, o en la mejora de los ya existentes. Como ya se ha expuesto con anterioridad, la complejidad del SVH es muy alta y en este sentido hay abiertas dos posibilidades en cuanto a su simulación. Por una parte se analizan en detalle las características visuales y se intenta reproducir su comportamiento mediante el uso de complejas funciones. Por otra, el SVH se trata como una caja negra, pretendiendo encontrar sistemas de medida alternativos que proporcionen valores cercanos a los subjetivos pero sin tener en cuenta la gran complejidad del SVH. El primero de los métodos es en el que se han basado la mayoría de las métricas, mientras que el segundo método es la base de medidas como el SSIM y, por tanto, el VSSIM. De forma general se puede relacionar el primer método con medidas basadas en detección de error y el segundo con medidas de detección de distorsión estructural.

Con todo esto se pretende dar una idea del gran campo de análisis y estudio del que se dispone a la hora de tratar con la subjetividad que plantea el SVH. Por ello, las ampliaciones que podrían hacerse para este proyecto pueden seguir diferentes líneas:

Una primera línea de investigación se centraría en aumentar la correlación con las valoraciones subjetivas. Aunque los resultados obtenidos son buenos, se ha comprobado que el algoritmo NTIA ofrece una correlación mayor. En este sentido es clave no sólo la

obtención de nuevos mecanismos que modelen mejor el SVH, sino la incorporación de módulos que simulen de forma más precisa los procesos de percepción a más alto nivel o la mejora de los existentes. En concreto un ámbito con grandes posibilidades es el estudio de la atención visual, relacionado con operaciones específicas de procesamiento de información. Dentro de estas operaciones específicas se encontraría la detección o selección de objetivos dentro de la escena, para ponderar la calidad de distintas zonas de forma más precisa que con los métodos de enmascaramiento actuales. Sin embargo, de nuevo se comprueba la dificultad de la tarea, pues hay que tener en cuenta aspectos tan complejos como los movimientos oculares.

Otro factor a tener en cuenta, muy relacionado con lo anterior, es el tiempo de ejecución. Se pretende conseguir unos resultados más precisos, pero también es deseable mantener una carga computacional no muy elevada. En efecto, sistemas de evaluación de calidad perceptual serían muy útiles en centros de producción de TV o en centros de *broadcast*, sobre todo actualmente con la incorporación de la televisión digital. En estos casos, la ejecución a tiempo real sería necesaria.

Por otro lado, como ya se ha comentado, es muy deseable desarrollar técnicas de medida que puedan evaluar la calidad sin referencia, esto es, sin disponer de la secuencia original. Sin embargo el desarrollo de estos algoritmos es extremadamente complejo (aunque los observadores humanos pueden evaluar sin dificultad la calidad de un vídeo sin ninguna otra referencia). En este sentido queda mucho trabajo por hacer, que se ha de basar en un análisis profundo y a más alto nivel de los sistemas de percepción humanos.

Por último se puede considerar la opción, no menos compleja de incluir el sonido a la hora de evaluar la calidad de las secuencias audiovisuales. Hay estudios [66] que demuestran que la percepción de calidad de una secuencia se ve modificada por la presencia de una señal sonora correspondiente a la escena en cuestión.

Anexo A

Presupuesto

A continuación se detalla el cálculo de los costes aproximados de realización del proyecto descrito en la presente memoria. El presupuesto está dividido en dos partes, por un lado los costes asociados a materiales empleados y por otro, los costes debidos a honorarios de las personas que han participado en el proyecto. Todo ello se une en un resumen en el que se contabilizan los gastos totales acumulados durante el periodo de desarrollo del proyecto.

A.1 Coste de material

El material empleado en la realización del proyecto consta de los siguientes elementos:

- *Equipo informático*: Se ha hecho uso de un ordenador personal cuyo valor aproximado es de 800€. Aunque el equipo se ha empleado exclusivamente en este proyecto, tras la finalización del mismo se aprovechará en otras labores, por lo que en concepto de amortización del material sólo se considerará el 25% de su precio, es decir, 200€.

- *Lugar de trabajo*, con las debidas condiciones de luz, calefacción, mantenimiento, más el mobiliario necesario; tiene un coste asociado de unos 900€/mes. Al tratarse de un espacio compartido por 3 personas, tendrá un coste individual asociado de 300€/mes. Puesto que el proyecto ha durado aproximadamente 8 meses, el coste asociado asciende a 2400€.

- *Material de oficina*: Dentro de este concepto se incluye todo el material desechable empleado en el proyecto, impresión de artículos, hojas, carpetas, etc. El total estimado es de 60€.

- *Coste de licencias software*: Para la realización de este proyecto se ha empleado el sistema operativo Windows XP Professional de Microsoft, con un coste por licencia de 220€, a amortizar en 4 años, por lo que en 8 meses sólo se consideran 37€ aplicables al proyecto. Para el desarrollo del software se ha empleado el programa Matlab de Mathworks, con un coste de licencia de 1950€, a amortizar en 4 años, por lo que el coste a cargar en el proyecto es de 325€.

- *Conexión a Internet*: La tarifa plana ADSL tiene un coste mensual de 41€. A lo largo de 8 meses, el coste será de 328€.

- *Café y bollería*: para los asistentes a la prueba subjetiva, con un coste aproximado de 40€.

- *Impresión, encuadernación y copias del informe:* El coste de la impresión y encuadernación de las copias necesarias del proyecto asciende a 150€.

En la siguiente tabla están resumidos los costes relacionados con el material.

Concepto	Coste	Cantidad	Total
Equipo informático	200 €	1 unidad	200 €
Lugar de trabajo	300 €/mes	8 meses	2400 €
Material de oficina	60 €	–	60 €
Costes de licencias	362 €	1 unidad	362 €
Conexión a Internet	41 €/mes	8 meses	328 €
Café y bollería	40 €	–	40 €
Impresión, copias, etc.	150 €	–	150 €
		TOTAL	**3540 €**

Tabla A.1 Gastos materiales

A.2 Coste de honorarios

La duración total del proyecto ha sido de ocho meses. De forma general se puede establecer un horario de trabajo de seis horas diarias. Teniendo en cuenta la semana laboral, de lunes a viernes, obtendríamos 30 horas semanales ó 120 horas mensuales. En total el

proyecto se traduciría en 960 horas de trabajo. En ese periodo se incluye el desarrollo de la memoria del proyecto.

Para los gastos personales se tendrán en cuenta los honorarios de un Ingeniero Técnico de Telecomunicación. Hasta hace poco, la Junta General del Colegio Oficial de Ingenieros Técnicos de Telecomunicación ofrecía una cantidad a modo de ejemplo para los libres ejercientes de la profesión, de forma que pudieran disponer de una referencia de los honorarios. Dicha cantidad definía unos honorarios de 65 €/hora. Hoy en día, el Ministerio de Economía y Hacienda ha remitido a todos los colegios profesionales una nota [64] en la que se indica que no se debe, ni siquiera, publicar un baremo con los honorarios ya que, éstos son libres y responden al libre acuerdo entre el profesional y el cliente.

Dada la situación, los honorarios deben definirse en función de una serie de factores: costes del ingeniero, desplazamientos, mecanografía, volumen de la actividad, etc. Teniendo en cuenta estos elementos y que, por lo general, un ingeniero técnico no suele cobrar menos de 40 €/hora, ni más de 70 €/hora, se definirán unos honorarios de 50 €/hora.

El salario del director de proyecto se estima, de forma general, como un 7% del coste total del proyecto. Dado que el coste, sin contar el salario del director, es de 51540 €, los honorarios de dirección ascienden a 3607,8 €.

A continuación se muestra la tabla que detalla el presupuesto de los gastos personales con los datos descritos anteriormente:

Concepto	Coste	Cantidad	Total
Ingeniero encargado del proyecto	50 €/hora	960 horas	48000 €
Director del proyecto	7%	(51540 €)	3607,80 €
		TOTAL (sin IVA)	**51607,80 €**
		IVA 16%	**8257,25 €**
		TOTAL	**59865,05 €**

Tabla A.2 Coste de honorarios

A.3 Total presupuesto

En total, teniendo en cuenta los gastos materiales y los gastos personales, obtenemos el presupuesto final:

Concepto	Coste
Coste de material	3540 €
Coste de honorarios	59865,05 €
TOTAL	**63405,05 €**

El presupuesto total de este proyecto asciende a SESENTA Y TRES MIL CUATROCIENTOS CINCO EUROS CON CINCO CÉNTIMOS.

Fdo: Carlos Esteban Baz Hormigos
Ingeniero Técnico de Telecomunicación especializado en Sonido e Imagen

Anexo B

Manual de Usuario

En este apartado se explicará brevemente el funcionamiento y manejo de la aplicación desarrollada para que cualquier usuario pueda utilizar el programa de forma correcta.

Para ejecutar la aplicación en una máquina concreta es necesario que el entorno Matlab (en su versión R2008a o superior) se encuentre instalado en dicha máquina. Existen dos formas de lanzar la aplicación: sin ejecutar Matlab, empleando el archivo ejecutable o desde la línea de comandos.

La primera forma, válida para entorno Windows, consiste simplemente en hacer doble clic en el archivo ejecutable `CalidadVideo`.

Fig. B.1 Archivo ejecutable

La segunda manera es válida para cualquier sistema operativo; consiste en ejecutar el programa desde la línea de comandos de Matlab indicando el nombre del archivo:

```
>>
>> CalidadVideo
```

De una u otra forma, al arrancar la aplicación aparecerá la pantalla mostrada en la figura B.2, en la que el usuario podrá configurar los parámetros de entrada. En ella se introducen las rutas absolutas de las secuencias original y distorsionada y se selecciona el parámetro *Rs* mediante el cursor deslizante.

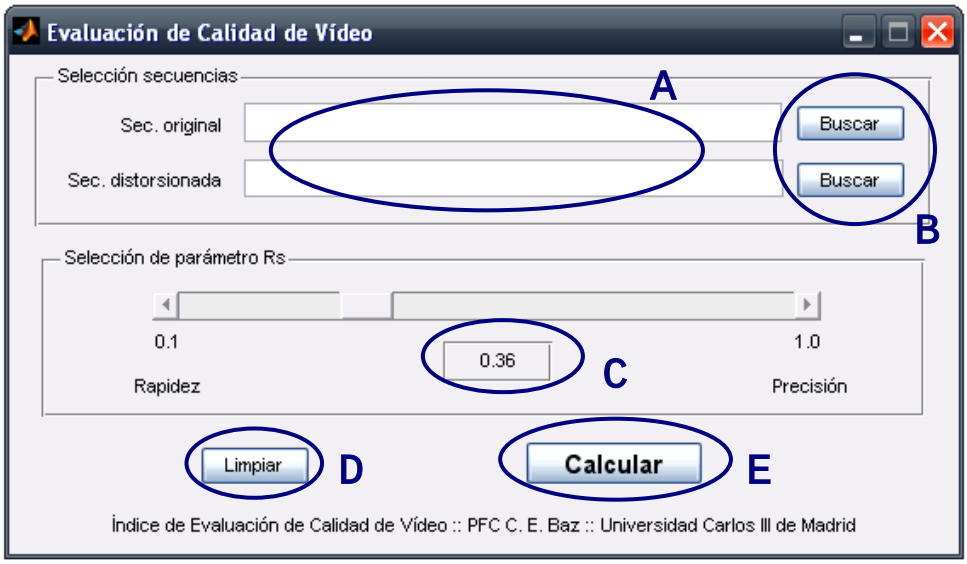


Fig. B.2 Descripción de la pantalla inicial

Las rutas se puede introducir de forma manual en los campos A, sin embargo al pulsar sobre los botones B, se abre el cuadro de búsqueda mostrado en la figura B.3 mediante el cual se halla fácilmente la ruta de cada una de las secuencias. Por otro lado, el cursor deslizante varía el parámetro *Rs*, cuyo valor actual se muestra en el cuadro C. El botón D limpia todos los valores introducidos y retorna la aplicación al estado inicial. Una vez que todos los parámetros son correctos, se continúa el proceso haciendo clic en el botón E.

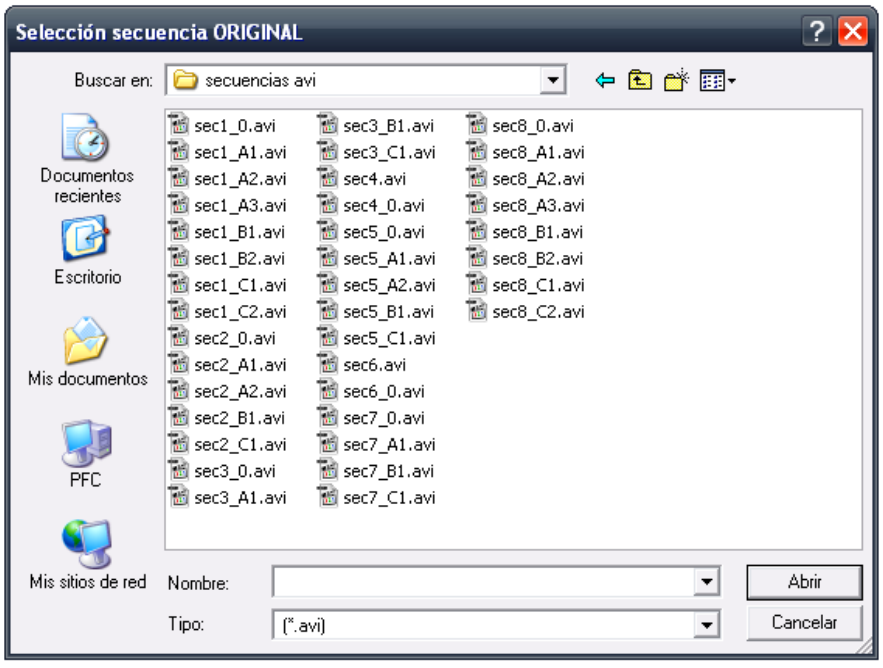


Fig. B.3 Cuadro de búsqueda de las secuencias

Si alguno de los campos de secuencia está vacío se mostraría el aviso de la figura B.4.

Fig. B.4 Aviso de campo de ruta distorsionada vacío

Si todo es correcto, comienza el cálculo, cargando previamente las secuencias (figura B.5). Durante la carga se pueden producir varios tipos de errores, mostrados en la figura B.6, tras los cuales se detendría la ejecución del programa.

Fig. B.5 Proceso de carga en memoria de secuencias

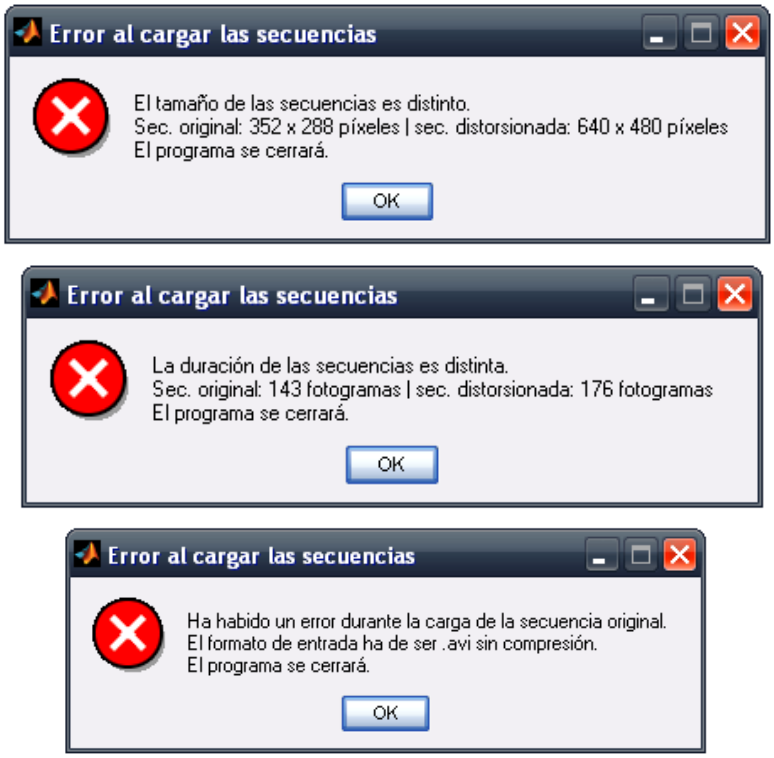


Fig. B.6 Mensajes de error de ejemplo posibles durante la carga de secuencias

Si todos los campos están completos y la carga de las secuencias se realiza de forma correcta, comienza el cálculo del índice. Durante el procesado se muestra la pantalla informativa de la figura B.7, que indica el porcentaje completado.

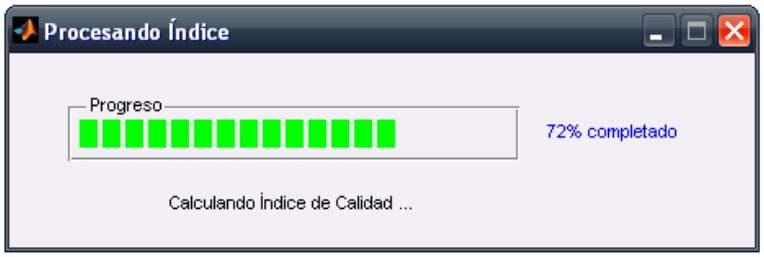


Fig. B.7 Pantalla de progreso del cálculo

Las etapas por las que se pasa en la fase de procesado son las siguientes:

- Carga de la secuencia original
- Carga de la secuencia procesada
- Cálculo del Índice de Calidad

De producirse, los errores mostrados en la figura B.6 ocurren tras la carga de ambas secuencias. Tras un procesado correcto, finalmente se muestra la pantalla de resultados (figura B.8).

En ella se puede observa la información proporcionada como resultado de la aplicación. En la zona A se indican parámetros informativos de las secuencias como nombres, duración en fotogramas y tamaño en píxeles. Los cuadros B y C muestran un fotograma de ejemplo de cada una de las secuencias, como comprobación de que se ha seleccionado la secuencia deseada y de que la aplicación la ha leído correctamente.

Finalmente, en la zona D se muestran los resultados de la medida de calidad. Se indica el índice obtenido (en escala unitaria y en escala subjetiva 1 a 5), el tiempo de ejecución necesario y el valor del parámetro *Rs* empleado. Tras ello, se muestra el gráfico E, que describe la evolución temporal del índice de calidad a lo largo de los fotogramas de la secuencia distorsionada.

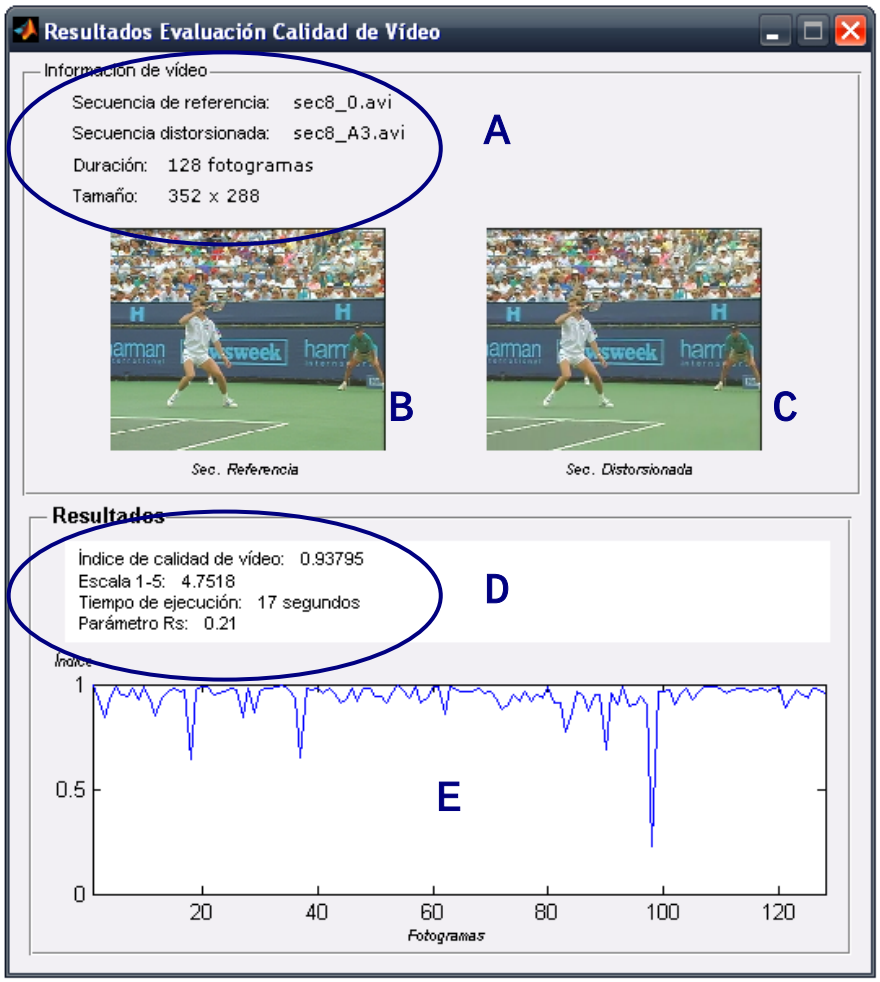


Fig. B.8 Descripción pantalla de resultados

Anexo C

Resultados detallados de la prueba subjetiva y de los algoritmos analizados

En este anexo se recogen de forma detallada todos los resultados obtenidos tanto en la prueba subjetiva como en la evaluación de los distintos algoritmos. Se incluyen en la memoria ya que esta información, principalmente la obtenida en el experimento subjetivo, puede ser útil en futuras investigaciones.

C.1 Prueba experimental

En las páginas siguientes se recogen las votaciones individuales de los 42 sujetos válidos que participaron en el experimento, así como el MOS y el intervalo de confianza de cada una de las secuencias.

En las dos tablas inferiores se indican algunos datos estadísticos de los sujetos participantes.

EDAD				
<21	21-25	26-30	31-35	>35
2	**15**	**14**	**7**	**4**

SEXO	
Hombres	**33**
Mujeres	**9**

Tablas C.1 Datos estadísticos de los observadores

		observador	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	secuencia																							
PS1	**sec1_B1**	entrenamiento	4	4	3	3	2	4	1	3	4	4	4	3	2	4	2	3	3	2	3	3	4	3
PS2	**sec8_A3**	entrenamiento	5	5	5	5	4	5	4	5	5	5	5	5	4	5	4	5	4	4	4	4	5	4
PS3	**sec4**	entrenamiento	5	4	2	4	4	4	3	5	4	5	3	4	5	3	5	5	3	4	3	5	4	3
PS4	**sec1_A1**		3	4	4	4	3	4	1	3	4	4	5	3	3	4	3	3	2	1	3	3	4	2
PS5	**sec3_C1**	repetición 1	3	4	3	2	3	5	2	3	4	4	2	4	3	3	3	4	3	2	3	4	3	2
PS6	**sec2_A1**		2	3	2	2	2	3	1	1	2	3	1	2	3	2	1	2	1	2	3	1	2	1
PS7	**sec2_A2**		4	3	2	2	2	3	1	1	3	3	2	2	2	3	1	3	1	3	3	1	2	1
PS8	**sec8_A1**		3	4	4	3	3	5	3	2	4	4	3	3	3	4	2	3	4	2	3	3	3	1
PS9	**sec7_B1**		3	3	3	3	3	3	2	2	4	4	5	2	3	3	2	3	2	3	3	2	2	1
PS10	**sec2_C1**		4	3	3	2	2	3	1	1	3	3	1	2	3	1	1	3	1	3	3	1	2	1
PS11	**sec1_C1**		3	4	5	4	4	3	3	2	4	4	4	3	3	4	2	2	3	3	4	2	2	2
PS12	**sec3_B1**		4	4	4	3	3	4	1	2	4	4	4	4	3	3	2	4	2	3	3	3	2	2
PS13	**sec5_B1**	repetición 2	4	4	5	4	4	5	3	2	4	4	5	4	4	5	4	4	4	4	3	4	3	3
PS14	**sec1_C2**	repetición 3	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	4	4	5	5	5	5	5
PS15	**sec6**		4	5	5	4	5	5	4	5	5	5	5	5	5	4	3	5	4	5	3	5	5	4
PS16	**sec7_A1**		4	4	4	3	4	2	2	2	3	3	4	3	4	4	2	3	2	3	3	3	3	2
PS17	**sec1_A3**		5	5	5	5	5	4	4	5	5	5	5	5	5	5	5	5	3	5	5	5	5	4
PS18	**sec8_B2**		5	5	5	5	5	5	4	5	5	5	5	5	5	5	4	4	4	4	5	5	5	5
PS19	**sec8_A3**		5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5
PS20	**sec1_A2**		5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5
PS21	**sec8_C1**		4	4	5	4	4	4	1	3	5	4	3	5	3	3	2	4	4	3	4	4	3	3
PS22	**sec5_A2**		5	4	5	4	4	5	3	4	5	4	4	4	4	4	4	4	4	4	3	4	4	4
PS23	**sec1_C2**	R3	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	4	4	4	5	5	5	5
PS24	**sec8_B1**		4	4	5	4	4	4	3	3	4	4	3	4	4	4	2	4	3	3	4	4	4	3
PS25	**sec7_C1**		4	4	3	3	3	4	1	3	3	3	3	3	3	4	2	4	1	3	3	2	2	2
PS26	**sec4**		5	4	3	4	4	5	2	3	5	5	3	4	5	3	5	5	4	5	3	5	3	3
PS27	**sec5_C1**		4	4	4	4	4	5	2	2	4	4	3	4	4	4	4	4	4	2	3	4	2	3
PS28	**sec2_B1**		3	3	3	2	3	4	1	2	2	2	2	3	2	2	2	3	1	2	3	2	2	2
PS29	**sec8_C2**		5	4	3	5	5	5	3	5	5	5	5	5	4	5	4	5	4	3	3	4	5	5
PS30	**sec5_B1**	R2	4	4	3	4	5	4	3	3	4	4	5	4	4	4	3	4	4	4	3	4	4	3
PS31	**sec3_C1**	R1	4	3	3	3	4	3	1	2	4	4	2	4	3	3	2	4	3	2	3	3	3	3
PS32	**sec1_B1**		3	3	4	3	3	3	2	2	3	3	4	4	2	3	2	3	3	1	3	2	2	2
PS33	**sec3_A1**		3	3	4	3	4	3	2	3	4	4	5	4	4	4	3	4	3	3	2	3	3	2
PS34	**sec8_A2**		5	4	5	4	5	5	3	5	4	5	5	5	5	3	4	5	4	4	5	4	5	4
PS35	**sec5_A1**		3	4	5	4	5	4	2	2	3	4	5	4	4	3	3	4	3	2	3	4	4	2
PS36	**sec1_B2**		5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	4	4	5	5	5	4

sec3_C1	Rep 1_med	3,5	3,5	3	2,5	3,5	4	1,5	2,5	4	4	2	4	3	3	2,5	4	3	2	3	3,5	3	2,5
sec5_B1	Rep 2_med	4	4	4	4	4,5	4,5	3	2,5	4	4	5	4	4	4,5	3,5	4	4	4	3	4	3,5	3
sec1_C2	Rep 3_med	5	5	5	5	4,5	5	5	5	5	5	5	5	5	5	4,5	4	4	4,5	5	5	5	5

edad	24	32	30	26	33	36	26	21	24	29	28	24	31	42	27	33	23	20	19	27	29	37
sexo	H	H	H	H	H	M	H	H	H	M	H	H	M	H	H	M	M	H	H	H	H	H
est. univ.	sí	sí	sí	no	no	no	sí	sí	sí	sí	sí	sí	sí	sí	sí	sí	no	sí	sí	sí	sí	sí

Tabla C.2 Tabla completa de resultados de la prueba subjetiva

23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	MOS	δ_{95}	Intervalo c. 95%
2	3	3	4	2	4	2	3	2	1	2	3	4	3	4	4	4	4	4	3			
5	4	4	4	5	4	5	4	5	3	4	4	5	5	5	5	4	4	4	4			
4	4	4	5	4	5	5	4	3	4	3	5	5	4	4	5	5	4	5	4			
3	3	2	3	1	3	2	4	2	1	2	2	4	4	4	4	3	4	4	3	**3,047619**	0,30209	**[3,05-0,30, 3,05+0,30]**
2	2	2	3	2	3	3	2	2	3	2	3	4	4	4	4	2	3	3	3			
1	1	1	2	1	2	1	2	1	1	1	2	3	2	2	2	1	2	3	1	**1,7619048**	0,21698	**[1,76-0,22, 1,76+0,22]**
1	2	1	3	1	3	1	3	1	2	2	3	3	3	2	4	1	2	3	2	**2,1666667**	0,27144	**[2,17-0,27, 2,17+0,27]**
2	3	2	2	2	2	3	3	2	2	2	3	3	4	3	4	2	2	4	2	**2,8809524**	0,25732	**[2,88-0,26, 2,88+0,26]**
1	2	1	2	2	2	2	1	2	3	2	3	4	4	3	3	2	3	3	2	**2,5714286**	0,26516	**[2,57-0,27, 2,57+0,27]**
2	2	2	3	1	3	1	2	2	1	2	2	3	3	2	4	2	3	3	1	**2,1666667**	0,27144	**[2,17-0,27, 2,17+0,27]**
3	3	2	3	2	3	2	3	3	1	2	2	4	4	4	3	2	4	4	3	**3,0238095**	0,26798	**[3,02-0,27, 3,02+0,27]**
4	4	3	3	2	3	1	2	3	3	2	3	4	4	4	4	2	4	4	2	**3,0714286**	0,27523	**[3,07-0,28, 3,07+0,28]**
3	4	3	4	3	4	4	3	2	4	4	3	5	5	4	5	4	4	4	3			
5	4	4	4	5	4	4	5	4	5	5	5	5	5	5	5	5	5	5	4			
4	5	4	4	4	4	5	3	4	5	4	4	5	5	5	5	4	5	5	4	**4,4761905**	0,18942	**[4,48-0,19, 4,48+0,19]**
3	4	2	3	1	3	2	1	3	3	2	3	4	2	3	4	3	4	3	3	**2,9285714**	0,25037	**[2,93-0,25, 2,93+0,25]**
5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	4	4	4	**4,7857143**	0,14055	**[4,79-0,14, 4,79+0,14]**
5	5	4	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	4	**4,8095238**	0,11875	**[4,81-0,12, 4,81+0,12]**
4	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	4	**4,8809524**	0,09794	**[4,88-0,10, 4,88+0,10]**
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	**4,952381**	0,06440	**[4,95-0,06, 4,95+0,06]**
3	4	3	3	4	3	3	3	2	3	3	3	3	4	4	4	3	4	4	3	**3,452381**	0,23979	**[3,45-0,24, 3,45+0,24]**
3	4	3	5	4	5	5	3	3	4	5	3	4	5	4	5	4	4	5	4	**4,0952381**	0,19588	**[4,10-0,20, 4,10+0,20]**
5	5	4	5	5	5	4	5	5	4	5	5	5	5	5	5	5	5	5	4			
3	4	3	3	3	3	3	4	4	4	4	3	4	4	4	4	4	3	4	3	**3,6190476**	0,17401	**[3,62-0,17, 3,62+0,17]**
2	3	2	3	1	3	3	2	2	3	2	4	3	4	3	4	2	3	3	3	**2,8095238**	0,24902	**[2,81-0,25, 2,81+0,25]**
4	4	3	5	4	5	4	3	2	5	4	5	5	4	5	4	5	4	5	4	**4,0952381**	0,27058	**[4,10-0,27, 4,10+0,27]**
3	3	3	4	4	4	3	2	3	4	3	4	3	5	4	5	4	3	5	2	**3,547619**	0,25732	**[3,55-0,26, 3,55+0,26]**
1	2	1	3	1	3	1	1	2	1	2	3	2	3	2	4	3	3	2	2	**2,2142857**	0,24280	**[2,21-0,24, 2,21+0,24]**
5	4	5	4	5	4	5	5	5	3	5	4	4	5	5	5	4	4	4	3	**4,4047619**	0,21947	**[4,40-0,22, 4,40+0,22]**
3	4	3	3	5	3	4	3	3	4	4	4	4	5	4	5	4	3	4	2			
3	4	2	3	2	3	2	3	3	3	3	3	2	4	4	4	3	4	4	3			
1	3	1	2	1	2	2	1	3	1	2	2	4	4	4	3	2	3	4	2	**2,547619**	0,28156	**[2,55-0,28, 2,55+0,28]**
4	3	2	3	1	3	2	2	3	4	3	3	3	4	4	4	3	3	4	3	**3,1904762**	0,24012	**[3,19-0,24, 3,19+0,24]**
4	4	5	3	4	3	5	5	4	4	5	4	4	5	4	5	3	4	5	3	**4,3095238**	0,21373	**[4,31-0,21, 4,31+0,21]**
3	4	3	3	2	3	2	2	3	4	3	3	3	4	4	5	4	3	4	2	**3,3571429**	0,27067	**[3,36-0,27, 3,36+0,27]**
5	4	5	5	5	5	4	5	5	4	5	4	5	5	5	4	4	4	5	4	**4,7142857**	0,13661	**[4,71-0,14, 4,71+0,14]**

2,5	3	2	3	2	3	2,5	2,5	2,5	3	2,5	3	3	4	4	4	2,5	3,5	3,5	3	**3,0119048**	0,19932	**[3,01-0,20, 3,01+0,20]**
3	4	3	3,5	4	3,5	4	3	2,5	4	4	3,5	4,5	5	4	5	4	3,5	4	2,5	**3,797619**	0,18973	**[3,80-0,19, 3,80+0,19]**
5	4,5	4	4,5	5	4,5	4	5	4,5	4,5	5	5	5	5	5	5	5	5	5	4	**4,7857143**	0,10509	**[4,79-0,11, 4,79+0,11]**

28	21	35	24	23	26	25	42	31	21	26	26	27	25	34	21	28	24	23	24
H	H	H	M	H	M	H	H	H	H	M	H	M	H	H	H	H	H	H	H
sí	sí	no	sí	sí	no	sí	no	sí	sí	sí	sí	sí	sí	sí	sí	sí	sí	sí	sí

C.2 Evaluación de algoritmos

A continuación se detallan las estimaciones de cada uno de los algoritmos analizados, así como de la medida propuesta para las secuencias empleadas en la prueba subjetiva. El valor indica la calidad de la secuencia en una escala 1 (muy mala) a 5 (perfecta).

Secuencias	Watson mod.	VSSIM	NTIA	Medida propuesta
sec1_A1	1,3723	4,8419	3,6256	4,0509
sec1_A2	3,6938	4,9657	4,8564	4,9064
sec1_A3	3,8024	4,9703	4,8439	4,9180
sec1_B1	1,2961	4,8346	3,5156	4,0361
sec1_B2	3,6733	4,9638	4,8308	4,8921
sec1_C1	1,0029	4,8384	3,5408	4,0407
sec1_C2	3,6052	4,9592	4,8272	4,8856
sec2_A1	1,0682	4,1864	2,3600	1,9037
sec2_A2	1,0464	4,3452	2,6756	2,4790
sec2_B1	1,0517	4,1684	2,2036	1,5397
sec2_C1	1,0285	4,2048	2,3928	1,9399
sec3_A1	1,0183	3,8646	2,9264	3,0836
sec3_B1	1,6919	3,8048	2,7952	3,0249
sec3_C1	1,9844	3,8037	2,6720	2,7414
sec4	1,1064	4,7123	4,2380	4,4256
sec5_A1	1,0358	4,0573	3,1596	3,2742
sec5_A2	1,6072	4,1747	3,3664	3,6311
sec5_B1	1,0469	4,0905	3,2648	3,3069
sec5_C1	1,1013	4,0926	3,2512	2,9639
sec6	1,0862	4,6772	4,1944	4,0959
sec7_A1	1,0129	4,3198	2,2496	1,4495
sec7_B1	1,1428	4,3308	2,2628	1,5212
sec7_C1	1,1395	4,3292	2,2668	1,1507
sec8_A1	1,1508	4,3915	3,6924	4,4397
sec8_A2	1,3462	4,8348	4,6716	4,7116
sec8_A3	1,4208	4,8713	4,7512	4,7696
sec8_B1	1,1365	4,3791	3,7472	3,4827
sec8_B2	1,5631	4,8398	4,6936	4,7135
sec8_C1	1,6615	4,2674	3,4612	3,2186
sec8_C2	1,8431	4,8141	4,6680	4,6531

Tabla C.3 Tabla completa de estimaciones de los distintos algoritmos

Por último, la siguiente tabla muestra las valoraciones subjetivas así como las estimaciones de los algoritmos para las secuencias empleadas en la prueba de formato D1.

Secuencias	MOS	Watson mod.	VSSIM	NTIA	Medida propuesta
src6_h264_64	1,4667	1,6869	3,7475	1,9504	1,5968
src6_h264_192	1,7333	2,3901	3,9095	2,3424	1,7031
src6_h264_256	2,0667	2,8803	3,9934	2,6256	1,7765
src6_h264_384	2,7333	3,3096	4,1083	3,0688	1,6413
src6_h264_512	3,2667	3,5490	4,1926	2,2964	2,2614
src6_h264_768	4,2000	3,8308	4,2994	3,5568	2,7231
src6_mas_K	3,4000	1,3121	4,7378	4,7668	3,4264
src5_h264_64	1,4000	1,7944	3,6484	1,4148	1,4155
src5_h264_384	1,6667	3,2367	4,2073	2,2792	1,7293
src5_h264_768	2,1333	4,0001	4,5159	3,0064	2,6947
src5_h264_1024	3,0667	4,2634	4,6139	3,3164	3,3491
src5_h264_2000	4,2000	4,7906	4,7956	4,0256	4,2204
src5_contraste	3,8667	1,5059	4,6172	4,6116	3,2578
src5_blur6	1,5333	2,3458	3,8782	1,7048	1,4916
src6_h264_64	1,4000	1,3493	2,7840	1,0460	1,2302
src2_h264_256	1,5333	1,3520	2,8406	1,3326	1,2417
src2_h264_384	2,1333	1,5139	3,5247	1,6172	1,3043
src2_h264_512	2,6667	2,2253	3,9727	2,1912	2,0866
src2_h264_768	3,8667	2,9730	4,2497	2,7488	3,1173
src9_h264_256	1,333	1,4528	3,7710	2,0532	1,4944
src9_h264_2000	4,2667	3,7520	4,6191	3,4468	3,8277

Tabla C.4 Tabla completa de estimaciones de los distintos algoritmos para las secuencias D1

Referencias bibliográficas

[1] ESKICIOGLU, A. M.; FISHER, P. S. "Image quality measures and their performance," *IEEE Trans. Communications*, vol. 43, pp. 2959-2965, diciembre 1995.

[2] GIROD, B. "What's wrong with mean-squared error," en *Digital Images and Human Vision*, A. B. Watson, pp. 207-220, MIT Press, 1993.

[3] WINKLER, S. "A perceputal distortion metric for digital color video" *SPIE*, vol. 3644, pp. 175-184, 1999.

[4] WANG, Z.; BOVIK A. C.; LU, L. "Why is image quality assessment so difficult?" *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, vol. 4, pp. 3313-3316, mayo 2002.

[5] LUKAS, F.J.; BUDRIKIS, Z.L. "Picture quality prediction based on a visual model" IEEE Vol. COM-30, pp.1679-1692, julio 1982

[6] VQEG. "Final report from the video quality experts group on the validation of objective models of video quality assessment", marzo 2000. Disponible en Web: *http://www.vqeg.org/* Ultimo acceso: 27/08/2009

[7] CORRIVEAU, P., *et al.*, "Video quality experts group: Current results and future directions" *SPIE Visual Comm. and Image Processing*, vol. 4067, junio 2000.

[8] GEISLER, W.S.; BANKS, M.S. "Visual performance," en *Handbook of Optics* (M. Bass, ed.), McGraw-Hill, 1995.

[9] WANDELL, B.A.; "Foundations of Vision", *Sinauer Associates, Inc*., 1995.

[10] CORMACK, L.K.; "Computational models, of early human vision" en *Handbook of Image and Video Processing* (A. Bovik, ed.), Academic Press, mayo 2000.

[11] HUBEL, D.H. "Ojo, cerebro y visión." *Servicio de Publicaciones, Universidad de Murcia,* 2000.

[12] PELI, E. "Contrast in complex images". *Journal of the Optical Society of America A, 2032-2040,* octubre 1990.

[13] DALY, S.; "The visible difference predictor: An algorithm for the assessment of image fidelity" en *Digital Images and Human Vision,* A. B. Watson, pp. 179-206. MIT Press, Cambridge, 1993.

[14] TAYLOR, C. "Image Quality Assessment via a Human Visual System Model", Ph.D. Dissertation, Purdue University, agosto 1998.

[15] PELI, E. "In search of a contrast metric: Matching the perceived contrast of gabor patches at different phases and bandwidths." *Vision Research*, 3217- 3224, 1997.

[16] NADENAU, M.J. "Integration of human color vision models into high quality image compression". *PhD thesis, Signal Processing, Laboratory, Swiss Federal Institute of Technology, Lausanne,* noviembre 2000.

[17] ARTIGAS, J.M. et al. "Óptica Fisiológica. Psicofísica de la Visión". Interamericana McGraw-Hill (ed.), 1995.

[18] OSBERGER, W. "Perceptual Vision Models for Picture Quality Assessment and Compression Applications". *Ph.D. dissertation*, Sch. Elect. Electron. Syst. Eng., Queensland Univ. Technol., Queensland, Australia, 1999.

[19] WATSON, A.B. "The cortex transform: rapid computation of simulated neural images" *Computer Vision, Graphics, and Image Processing*, vol. 39, pp. 311-327, 1987.

[20] LUBIN, J. "The use of psychophysical data and models in the analysis of display system performance," en *Digital Images and Human Vision,* A. B. Watson, pp. 163-178, Cambridge, Massachusetts: The MIT Press, 1993.

[21] LUBIN, J. "A visual discrimination model for image system design and evaluation," en *Visual Models for Target Detection and Recognition,* E. Peli, pp. 207-220, Singapore: World Scientific Publisher, 1995.

[22] TEO P.C.; HEEGER, D.J. "Perceptual image distortion" en *Proc. IEEE Int. Conf. Image Processing*, pp. 982-986, 1994.

[23] HEEGER D.J.; TEO P.C., "A model of perceptual image fidelity" en *Proc. IEEE Int. Conf. Image Proc.*, pp. 343-345, 1995.

[24] WATSON A.B.; SOLOMON J.A., "Model of visual contrast gain control and pattern masking" *Journal of Optical Society of America*, vol. 14, no. 9, pp. 2379-2391, 1997.

[25] XU, W.; HAUSKE, G. "Picture quality evaluation based on error segmentation", *SPIE, vol. 2308, pp. 1454-1465, 1994.*

[26] OSBERGER, W.; BERGMANN, N.; MAEDER, A. "An automatic image quality assessment technique incorporating high level perceptual factors" en *IEEE Int. Conf. Image*, pp. 414-418, 1998.

[27] DALY, S. "The visible difference predictor: An algorithm for the assessment of image fidelity," en *Proc. SPIE*, vol. 1616, pp. 2-15, 1992.

[28] BURT P.J.; ADELSON, E.H. "The Laplacian pyramid as a compact image code" *IEEE Trans. Communications*, vol. 31, pp. 532-540, abril 1983.

[29] FREEMAN W.T.; ADELSON, E.H. "The design and use of steerable filters" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, pp. 891-906. 1991.

[30] SIMONCELLI E.P., *et al.* "Shiftable multi-scale transforms" *IEEE Trans. Information Theory*, vol. 38, pp. 587-607, 1992.

[31] WATSON, A.B., "DCTune: A technique for visual optimization of DCT quantization matrices for individual images" en *Society for Information Display Digest of Technical Papers*, vol. XXIV, pp. 946-949, 1993.

[32] SAFRANEK R.J.; JOHNSTON, J.D. "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression" en *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Processing*, pp. 1945-1948, mayo 1989.

[33] WOODS J.W.; O'NEIL, S.D. "Subband coding of images" *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, pp. 1278-1288, octubre 1986.

[34] BRADLEY, A.P. "A wavelet difference predictor" *IEEE Trans. Image Processing*, vol. 5, pp. 717-730, mayo 1999.

[35] ANTONINI, M., *et al.* "Image coding using the wavelet transform" *IEEE Trans. Image Processing*, vol. 1, pp. 205-220, abril 1992.

[36] LAI Y.K.; KUO, C.J. "A Haar wavelet approach to compressed image quality measurement" *Journal of Visual Communication and Image Understanding,* vol. 11, pp. 17-40, marzo 2000.

[37] DAMERA-VENKATA, N., *et al.* "Image quality assessment based on a degradation model" *IEEE Trans. Image Processing*, vol. 4, pp. 636-650, abril 2000.

[38] KARUNASEKERA S.A.; KINGSBURY, N.G. "A distortion measure for blocking artifacts in images based on human visual sensitivity" IEEE Trans. Image Processing, vol. 4, pp. 713-724, junio 1995.

[39] KARUNASEKERA S.A.; KINGSBURY, N.G. "A distortion measure for image artifacts based on human visual sensitivity," en Proc. IEEE Int. *Conf. Acoust., Speech, and Signal Processing*, vol. 5, pp. 117-120, 1994.

[40] CHOU C.H.; LI, Y.C. "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 5, pp. 467-476, diciembre 1995.

[41] MIYAHARA, M.; KOTANI, K.; ALGAZI, V.R. "Objective picture quality scale (PQS) for image coding" *IEEE Trans. Communications*, vol. 46, pp. 1215-1225, septiembre 1998.

[42] TAN, K.T.; GHANBARI, M.; PEARSON, D.E. "An objective measurement tool for MPEG video quality" *Signal Processing*, vol. 70, pp. 279-294, noviembre 1998.

[43] VAN DEN BRANDEN, C.J. "Perceptual models and architectures for video coding applications", *PhD thesis*, Swiss Federal Institute of Technology, agosto 1996.

[44] VAN DEN BRANDEN, C.J. "A working spatio-temporal model of the human visual system for image restoration and quality assessment applications" en *IEEE Int. Conf. Acoust.*, Speech, and Signal Processing, pp. 2291-2294, 1996.

[45] VAN DEN BRANDEN, C.J., *et al.* "Quality assessment of motion rendition in video coding" *IEEE Trans. Circuits and Systems for Video Tech*., vol. 9, pp. 766-782. agosto 1999.

[46] WATSON, A. B., *et al.* "DVQ: A digital video quality metric based on human vision" *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 20-29, 2001.

[47] WATSON, A.B. "Toward a perceptual video quality metric" en *Proc. SPIE Human Vision and Electronic Imaging III*, vol. 3299, pp. 139-147, enero 1998.

[48] TAN K.T.; GHANBARI, M. "A multi-metric objective picture-quality measurement model for MPEG video" *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 10, pp. 1208-1213, octubre 2000.

[49] YU, Z., *et al.* "Vision-model-based impairment metric to evaluate blocking artifact in digital video," *Proceedings of the IEEE*, vol. 90, pp. 154-169, enero 2002.

[50] WANG Z.; BOVIK, A.C. "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81-84, marzo 2002.

[51] WANG, Z.; LU L.; BOVIK, A.C. "Video quality assessment using structural distortion measurement," *Proc. IEEE Int. Conf. Image Proc.*, septiembre 2002.

[52] *Preliminary Draft New Recommendation* "Objective perceptual video quality measurement techniques for digital broadcast television in the presence of a full reference" Recomendación de la ITU-R.

[53] *Draft Revised Recommendation J.144* "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference" Recomendación de la ITU-R.

[54] ROBERTSON M.A.; STEVENSON, R.L. "DCT Quantization Noise in Compressed Images" *IEEE Transactions on Circuits and Systems For Video Technology*, Vol. 15, No. 1, enero 2005

[55] WU H.R.; RAO, K.R. "Digital Video Image Quality and Perceptual Coding", 2006, CRC Press

[56] PINSON M.H.; WOLF, S. "A new standarized method for objectively measuring video quality", noviembre 2003

[57] XIAO F. "DCT-based Video Quality Evaluation" Proyecto final para EE392J, 2000

[58] WANG, Z.; LU L.; BOVIK A.C. "Video quality assessment based on structural distortion measurement" Signal Processing: Image Communication, vol. 19, febrero 2004

[59] BHAT, A.; RICHARDSON I; KANNANGARA S. "A novel perceptual quality metric for video compression" Robert Gordon University

[60] *Recomendación ITU-T P.910,* "Subjective video quality assessment methods for multimedia applications", Abril 2008

[61] *Recomendación ITU-R BT.500-11,* "Methodology for the subjective assessment of the quality of television pictures", Junio 2002

[62] Software MSU Video Quality Measurement Tool. Último acceso 9/09/2009. http://compression.ru/video/quality_measure/video_measurement_tool_en.html

[63] MSU Graphics & Media Lab de la Universidad Estatal de Moscú. Último acceso 9/09/2009. http://graphics.cs.msu.ru/

[64] COITT. Último acceso: 4/09/2009
http://www.coitt.es/res/libredocs/Honorarios.pdf

[65] WANG, Z.; SHEIKH H.R.; BOVIK A.C. "Objective Video Quality Assesment" capítulo 41 en *The Handbook of Video Databases: Design and applications,* septiembre, 2003

[66] FRATER, M.R.; ARNOLD, J.F.; VAHEDIAN, A. "Impact of audio on subjective assessment of video quality in videoconferencing applications" *IEEE Journal of Selected Areas in Comm.*, vol. 11, pp. 1059-1062, septiembre, 2001.