UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE INGENIERÍA

Evaluación subjetiva de la calidad de video en 4K

MEMORIA DE PROYECTO PRESENTADA A LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD DE LA REPÚBLICA POR

ALEJANDRO GODAY Y LAURA GOMEZ

EN CUMPLIMIENTO PARCIAL DE LOS REQUERIMIENTOS PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO ELECTRICISTA.

TUTORES

RAFAEL SOTELO	UNIVERSIDAD DE LA REPÚBLICA
JOSÉ JOSKOWICZ	UNIVERSIDAD DE LA REPÚBLICA

TRIBUNAL

FEDERICO LECUMBERRY	UNIVERSIDAD DE LA REPÚBLICA
JUAN PABLO GARELLA	UNIVERSIDAD DE LA REPÚBLICA
JOSÉ JOSKOWICZ	UNIVERSIDAD DE LA REPÚBLICA
RAFAEL SOTELO	UNIVERSIDAD DE LA REPÚBLICA

Agradecimientos

A quienes participaron de las pruebas: Carlos Goday, Susana Ruiz, Lilián Giglio, Leonardo Gomez, Laura Tipoldi, Felipe, Antonella, Matías Cabrera, Emiliana Botto, Matías Bonora, Santiago Topham, Ariel Gomez, Alexis Rodríguez, Patricia García, Carlos Páramos, Bruno Bianchi, Lorgio Riva, Eduardo Scaraffoni, Pablo Balliva, Sylvia Díaz, Ignacio Zas, Nicolás Quartino, Daniel Lindner, Santiago García, Alibet León Ramos, Sebastián Sequeira, Alfonso Michelin, Sergio Tambucho, Santiago Silva Dalla, Felipe Barbeito, Manuela Cedres, Manuela Viola, Elena Kremer, Alexander Ocaño, Eduardo Viglietti, Alejandro Bonjour, Lucia Rodríguez, Silvia Llambi, Diego Vallejo, Paola Abefasi, Jessica Silva, Analí Mas, Giuliano Mancuso, Sebastian Blanco, Mariana Riera, Maria Cserni, Nicolás Morena, Victoria Roses, Gabriela Florin, Lucía Romeo, Fernando Albano, Mariana Goycoechea, Federico Gonzalez, Debora Bittencourt, Gerardo Rego, Esteban Goñi, Fabiana Francia, Yuniel Carcases Borges, Veronica Ruiz, Juan Bence, Andrea Brandon, María Gómez, Luciano Santos, Adriana Casteran, Rodolfo Rossado, Juan Espinosa, Victoria Mallea, Matias Aguirrezabal, Martín Riffaud, Alejandra Pérez, Milton Bentos, Karl Andreas Hennig, Victoria Rincon, María Misa, Alejandro Canabé, Nicolás Barlocco, Germán Abad Njerš, Mariana Ingold, Alexis Zamas, Carlos Turri, Rubén Rozenweig, Mario Testore, Carlos Galucci, Andrea Lasa, Carlos Dorrego, Fabiana Mori, Laura Bonilla, Cesar Braganca, Tomás Dorrego, Javier Sarasua, Álvaro Ferro, Romina Zepedeo, Fernando Caamaño, Michel Rivas, Beatriz Viglietti, Álvaro Viscardi, Kenny Diaz, Flavio Bentancour, Danilo Zimermann, Sebastián Vieira, Maximiliano Dutra, María Goday, Luís García, Sebastián Peña, Eduardo Freijanes, Nicolás Mallea, Doreli Perez, Ivette Stadler y Maximiliano Pérez.

A Fernando Angeloro, quien nos recibió en el laboratorio de televisión digital en el LATU.

Al Instituto de Ingeniería Eléctrica por cedernos la sala para realizar las pruebas.

A DINATEL por el préstamo de los equipos al Instituto de Ingeniería Eléctrica de la Facultad de Ingeniería.

Resumen

En este proyecto se diseñó y empleó una metodología para realizar evaluaciones subjetivas de calidad de video en 4K con siete personas. Esto se hizo con el fin de analizar, en general, la influencia que la posición de un observador tiene sobre las calificaciones que asigna al contenido que se le muestra, y en particular, la viabilidad de hacer una prueba con esta cantidad de personas.

La implementación de dicha metodología se llevó a cabo a lo largo de dieciséis sesiones. De estas dieciséis sesiones de evaluación subjetiva, una de ellas fue realizada el 25 de octubre de 2018 en el salón 6 del edificio Los Talas del LATU, mientras que las quince sesiones restantes fueron realizadas entre el 22 de noviembre y el 14 de diciembre de 2018 en la sala IEEE del Instituto de Ingeniería Eléctrica de la Facultad de Ingeniería de la Universidad de la República.

Esta documentación se divide en cinco capítulos, cuyo contenido se explica a continuación:

- Capítulo 1: Introducción
  Explica a grandes rasgos el contexto en el cual cobra importancia la realización de pruebas subjetivas, las recomendaciones de la Unión Internacional de Telecomunicaciones que son de utilidad al momento de hacer evaluaciones subjetivas en calidad de video, y las limitaciones de estas recomendaciones que motivan parcialmente la realización de este proyecto. Se explican en este capítulo también el tema a tratar en nuestro proyecto, su motivación y su objetivo.

- Capítulo 2: Definición de la sesión de prueba
  En este capítulo se describen todos aquellos aspectos relacionados a la planificación de las sesiones de evaluación de calidad realizadas y al análisis estadístico posterior. Entre los aspectos previamente mencionados se incluyen: Las pruebas visuales realizadas a los observadores, el acondicionamiento lumínico de las salas, los equipos utilizados, la distribución espacial de los observadores, el método de selección y compresión de contenido, la metodología de evaluación subjetiva utilizada, la aplicación usada para recibir las votaciones y los datos de los observadores junto con los cambios que se le introdujeron para este proyecto, y la matemática involucrada en el análisis de los resultados expuestos a posteriori.

- Capítulo 3: Sesiones de evaluación
  Se presentan en este capítulo detalles de las sesiones realizadas y una caracterización demográfica de los participantes de nuestras sesiones a partir de los datos que ellos proveyeron. Se explican también algunos aspectos de la sesión que no podían planificarse, como por ejemplo: El comportamiento de los observadores, inconvenientes técnicos, opiniones generales de los observadores acerca de la sesión, etc.

- Capítulo 4: Análisis de resultados
  Se detallan y analizan los resultados obtenidos a partir de las votaciones provistas por los observadores a través de la aplicación utilizada, utilizando las herramientas matemáticas descriptas en el capítulo 2.

- Capítulo 5: Conclusiones y trabajo futuro
  Se presentan las conclusiones finales obtenidas a partir del trabajo realizado a lo largo de este proyecto. Además, se proveen líneas de acción para trabajo futuro.

Nota aclaratoria

El presente trabajo se basa en los resultados de sesiones de evaluación de calidad de video en 4K realizadas por Laura Gómez, Alejandro Goday y Federico Páramos, como parte de la asignatura Proyecto de Ingeniería Eléctrica, entre el 3 de marzo de 2018 y 14 de diciembre de 2018. Podría existir otro documento basado en los mismos datos escrito por Federico Páramos.

Tabla de contenido

1 Capítulo 1: Introducción

1.1 Contexto

Con la llegada de la resolución de video 4K (3840 x 2160 píxeles), cuatro veces mayor a su predecesora HD (1920 x 1080), y del códec de video de alta eficiencia HEVC, dos veces más eficiente que su predecesor AVC en términos de calidad en función de tasa de bits [1], resulta natural la idea de hacer nuevas codificaciones de video con resolución 4K utilizando HEVC.

A su vez, junto al incremento en el uso de servicios de video a demanda, los consumidores de estos servicios buscan cada vez mejor calidad de video. Como resultado, la evaluación subjetiva de calidad de video, que es el proceso de emplear observadores humanos para calificar la calidad de video basándose en percepción individual, se ha convertido en un tema de alto interés, tanto en la academia como en la industria. En particular, resulta de interés la evaluación subjetiva de calidad de video, para videos con resolución 4K codificados con HEVC, dado que estos son adelantos que se buscan introducir tanto al mercado de video a demanda como al de televisión por cable.

Dado que se necesita evaluar calidad de video, deben tomarse una serie de decisiones acerca de cómo dichas evaluaciones se llevarán a cabo. A estos efectos existen documentos creados por la Unión Internacional de Telecomunicaciones (ITU, sigla en inglés de *International Telecommunications Union*) llamados "recomendaciones", uno de cuyos fines es proveer lineamientos para la realización de evaluaciones de calidad de video en ambientes de laboratorio. Dichos lineamientos usualmente refieren a las condiciones de la sala (por ejemplo: iluminación, distancia entre un observador y la pantalla), la cantidad de observadores a reclutar para la evaluación dependiendo del fin, los métodos de calificación de calidad de video, y la forma de procesar y presentar los resultados de la evaluación.

Sin embargo, las recomendaciones existentes proveen lineamientos para evaluaciones con videos de definición estándar (SD, del inglés *Standard Definition*) o de alta definición (HD, del inglés *High Definition*), pero no para evaluaciones con videos de ultra alta definición (o si se quiere, resolución 4K). Esto nos lleva a intentar definir nuevos lineamientos para evaluaciones con video de esta resolución, que pudieran ser parcialmente basados en los existentes para HD y SD.

En el contexto de la realización de evaluaciones subjetivas de calidad con videos en resolución SD y HD, las recomendaciones de la ITU de mayor relevancia resultan ser la ITU-R BT.500-13 [2], donde se provee una metodología para la evaluación subjetiva de calidad de imágenes en televisión, la ITU-T REC P.910 [3], donde se proveen métodos de evaluación subjetiva para aplicaciones multimedios, y finalmente la ITU-T REC P.913 [4], que se enfoca en la evaluación de calidad audiovisual en internet y en televisión.

1.2 Temas a tratar

Este proyecto trata con la calidad de experiencia en la visualización de videos 4K en televisores de dicha resolución. En particular se enfoca en la valoración subjetiva de contenidos visuales por parte de observadores inexpertos, y en la forma de hacer sesiones de evaluación subjetiva para obtener estas valoraciones. El factor principal a tratar en este proyecto, en lo que refiere a la valoración subjetiva, será la posición del usuario respecto del televisor.

1.3 Motivación

Usualmente, en estudios con evaluaciones subjetivas de calidad en 4K, las sesiones de evaluación se llevan a cabo con uno o dos sujetos evaluadores (a los cuales referiremos en este texto como *observadores*). Esto permite que los observadores sean posicionados próximos al televisor y a la misma distancia. Dependiendo de la prueba, puede llegar a ser necesario reclutar decenas, o incluso cientos de observadores para poder obtener resultados estadísticamente significativos. Esto implicaría un elevado número de sesiones si se realizan con una o dos personas. Lo anterior presenta inconvenientes de varios tipos, como, por ejemplo: de coordinación con los observadores, de disponibilidad de un ambiente de laboratorio, de tiempo, y de dinero por gastos del ambiente a utilizar. Sería una ventaja entonces poder reducir a un mínimo la cantidad de sesiones de evaluación, aumentando la cantidad de observadores en cada sesión.

1.4 Objetivo

Teniendo en mente las consideraciones anteriores, se desea determinar si es viable realizar sesiones de evaluación subjetiva de calidad de video en 4K con siete personas. Es decir, si las puntuaciones reportadas para las diferentes posiciones no presentan diferencias estadísticamente significativas respecto a las de la posición central de visión. En caso de que alguna ellas presentasen diferencias estadísticamente significativas con la óptima, se descartaría dicha posición como viable.

2 Capítulo 2: Definición de la Prueba

El primer paso para realizar las pruebas subjetivas de calidad de video en 4K fue elaborar el plan de pruebas. En ese documento se definen las condiciones en que se realizará la prueba: acondicionamiento de la sala, datos y características de los observadores, equipamiento a usar, los videos a utilizar, el método de evaluación y el posterior procesamiento de los datos.

El cometido de este capítulo es presentar los puntos detallados en el plan de pruebas.

2.1 Objetivo

Determinar si es viable realizar sesiones de evaluación subjetiva de calidad de video en resolución 4K (3840 x 2160) con siete observadores. En particular, se quiere determinar si al hacer sesiones con siete observadores, la posición de cada observador es un factor de influencia en su valoración subjetiva del contenido que se le muestra.

2.2 Parámetros

2.2.1 Observadores

Para que una persona pueda participar como observador, no puede ser experto ni en calidad ni en procesamiento de video, deben hacérsele pruebas para determinar su nivel de agudeza visual y si es o no daltónico, y deben tomársele una serie de datos (especificados más adelante).

- **Prueba de agudeza visual**
  Esta prueba se realiza con la gráfica de Snellen (véase anexo A.1). Una persona tomándolo se para a 2.8 metros de distancia de la gráfica, y lee en voz alta las letras que se le señalan de cada fila, comenzando por la fila de más arriba. La fila más pequeña que el individuo puede leer, indica su nivel de agudeza visual. Aquellos individuos que acostumbran usar lentes, los usan durante la realización de esta prueba.

- **Prueba de daltonismo**
  Esta prueba es basada en el test de color Ishihara, utilizado para detectar daltonismo. Éste involucra 24 (o 38) placas pseudo-isocromáticas. Cada una de ellas muestra o bien un número, o algunas líneas. Los números que se ven son declarados, y cada respuesta debería darse sin más de tres segundos de retraso. En un examen a gran escala, el test puede ser simplificado al uso de solamente seis placas (1, 2, 7, 9, 12 y 14) que pueden verse en el anexo A.2. Para las pruebas subjetivas se utiliza este test abreviado de seis placas, donde se considerarán daltónicos a aquellos participantes que no puedan ver correctamente más de una placa. Puede ser necesario variar el orden si se sospecha que el sujeto puede intentar engañar de forma deliberada.

- **Datos requeridos**
  A todos los observadores se les toman los siguientes datos:
  - Cédula de identidad.
  - Posición (número de asiento)
  - Edad.
  - Sexo (Hombre/Mujer).
  - Nivel educativo.
  - Cantidad de horas que mira contenidos audiovisuales en su televisor/tablet/celular/PC por día.

Al momento de realizar la prueba, los observadores reciben instrucciones sobre la misma. Los detalles e instrucciones se dan de forma oral y se detallan en el anexo B.

2.2.2 Acondicionamiento de la sala

La primera sesión de evaluación subjetiva de calidad de video en 4K se realizó en el Laboratorio de Televisión Digital del LATU, mientras que las restantes quince se realizaron en la sala IEEE del Instituto de Ingeniería Eléctrica (los detalles se presentan en el capítulo 3). El acondicionamiento del espacio a utilizar, sigue las pautas vistas en la recomendación ITU-R BT.500-13 [2]. Para iluminar la sala se usó únicamente la luz de un foco, habiendo bloqueado toda entrada de luz a la sala. Para realizar las medidas de luminancia y temperatura de color se utilizó un colorímetro. Detalles de estos equipos pueden verse en la sección 2.2.3.

En la Tabla 1 se listan las medidas de los valores de luminancia mencionados en ITU-R BT.500-13 [2].

Valor	Medida LATU (**lux**)	Medida IIE (**lux**)
Luminancia de pantalla inactiva.	3,82	3,82
Cresta de luminancia.	173	234,8
Luminancia de la pantalla cuando sólo se muestra el nivel del negro en una sala completamente oscura.	3,54	1,36
Luminancia del blanco más intenso en sala completamente oscura.	162,6	224,6
Luminancia del fondo detrás del receptor.	4,14	8,96

Tabla 1: Valores de luminancia medidos

En la Tabla 2, se contrastan las pautas de la recomendación con las medidas en cada sala.

Parámetros	Recomendación ITU-R BT.500-13	Medida LATU	Medida IIE
Relación entre la luminancia de pantalla inactiva y el valor de cresta de la luminancia.	≤0,02	0,022	0,016
Relación entre la luminancia de la pantalla, cuando sólo se muestra el nivel del negro en una sala completamente oscura, y la correspondiente al blanco más intenso.	≈0,01	0,02	0.0061
Brillo y contraste de la imagen.	Establecido vía PLUGE	Valores de fábrica.	Valores de fábrica.
Ángulo máximo de observación con respecto a la normal (este valor se aplica a las pantallas de tubo de rayos catódicos (TRC), para otro tipo de pantallas se están estudiando los valores adecuados)	30°	40°	40°
Relación entre la luminancia de fondo detrás del receptor de imágenes y el valor de cresta de luminancia de la imagen.	≈0,15	0,024	0,038
Cromaticidad del fondo.	D_{65}	5457 K	4304 K
Otra iluminación de la sala	Débil	Débil	Débil

Tabla 2: Parámetros de acondicionamiento de la sala

En la Tabla 2 puede verse que el ángulo máximo de observación es mayor al definido en la recomendación. En la sección 2.2.4 puede verse la razón de esta elección.

En cuanto a la iluminación, solo contábamos con un foco que se colocó en una esquina de la sala buscando el confort visual de los participantes. Se trató de calibrar el equipo de forma de obtener valores lo más cercanos a la recomendación posible. Al margen de las diferencias respecto a la recomendación, la iluminación del ambiente no parecía incidir negativamente sobre el proceso de calificación de los videos a nuestro criterio.

2.2.3 Especificaciones de los equipos utilizados

2.2.3.1 Televisor

Especificaciones del Televisor	
Pantalla	Cristal Líquido
Marca	Sony Bravia
Diagonal	55 pulgadas
Número de Modelo	***XBR-55X705D***
Resolución Nativa	3840 x 2160 (4K)
Altura de Sección Visible (***H***)	0,68 metros

2.2.3.2 Computadora

Especificaciones de la Computadora	
Marca	Acer
Modelo	Predator 17 G9-792-73UG
Sistema Operativo	Windows 10 Home
Procesador	Intel Core i7-6700HQ
Pantalla	Panel LCD de 17 pulgadas UHD 4K
Tarjeta Gráfica	NVIDIA GeForce GTX 980M
Memoria	32GB DDR4
Almacenamiento	512GB+512GB RAID +1000GB HDD
WLAN/Bluetooth	802.11 ac +BT
Batería	Li-on de 8 celdas

2.2.3.3 Router

Especificaciones del Router	
Marca	TP-Link
Modelo	TL-WR840N
Tipo	Inalámbrico N 300Mbps
Alimentación	9V-0.6A

2.2.3.4 Foco

Especificaciones del Foco	
Marca	Litepanels
Modelo	ASTRA 1X1 EP BI-COLOR
Panel	LED
Vin	13-24VDC

2.2.3.5 Colorímetro

Especificaciones del Colorímetro	
Marca	Sekonic
Modelo	C-700R

2.2.3.6 Móviles de prueba

Especificación de los móviles de prueba	
Modelo	**Cantidad**
Samsung GALAXY Mini	3
Samsung GALAXY S6 EDGE PLUS	2
Samsung GALAXY J1 ace	1
Samsung GALAXY CORE LTE Prime	1

Todos los móviles cuentan con sistema operativo Android.

2.2.4 Distribución de los observadores

La distribución espacial de los observadores es la de la Figura 1.

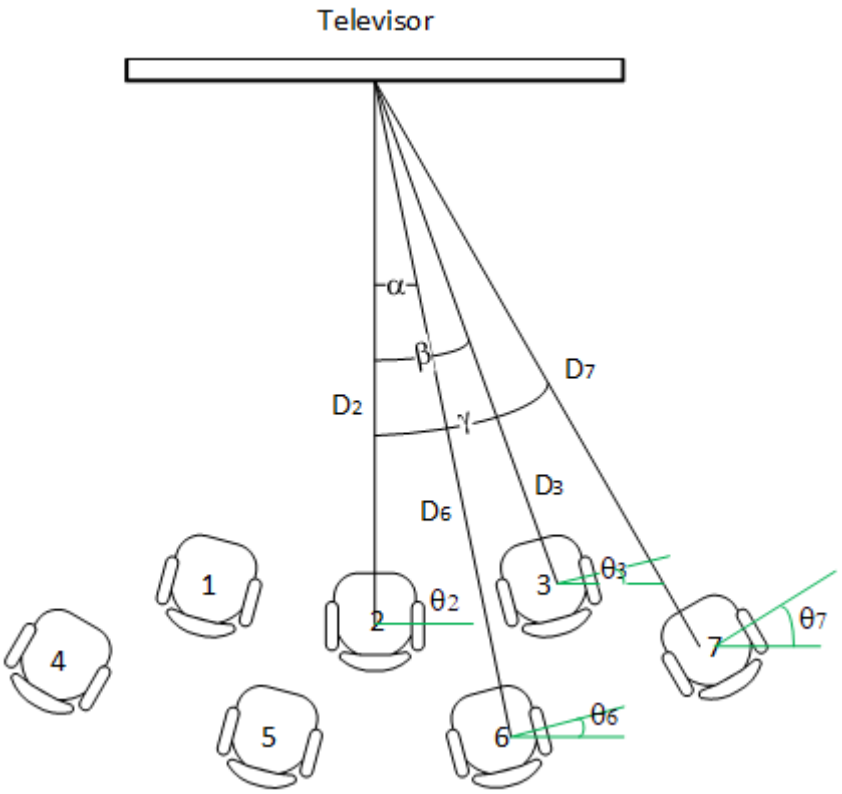


Figura 1: Disposición de los asientos

Al momento de elegir la disposición geométrica de los asientos fue necesario tomar en cuenta las siguientes restricciones:

1. Cada asiento debe estar lo suficientemente lejos del televisor como para que un observador pueda observarlo en su totalidad sin incomodidad.
2. Cada asiento debe estar lo suficientemente cerca como para poder analizar en detalle la imagen, distinguiendo, por ejemplo, efectos de cuadriculado de la imagen debido a pérdidas de compresión del video.
3. Los asientos de adelante (1, 2 y 3) deben estar a una distancia tal de los asientos de atrás (4, 5, 6 y 7), que los observadores puedan sentarse sin apretarse.
4. Ninguna persona sentada en un asiento de adelante debe obstruir la visión de ninguna persona de atrás.
5. En lo posible se debe tratar de respetar la recomendación ITU-R BT.500-13 [2] en lo que refiere al ángulo de visión. O sea, debe intentarse que los ángulos de visión (referidos por α, β y γ) sean menores a 30°.

A efectos de contemplar la restricción 1, se decidió colocar el asiento delantero central (número 2 en la Figura 1) a una distancia de aproximadamente dos veces la altura visible de la pantalla (2***H***). Si bien en las recomendaciones no se definen distancias preferidas de visión para 4K, se eligió 2H en base a trabajos previos, como por ejemplo [5], [6], [7], [1], [8], [9] y [10]. En estos trabajos se manejan distancias de 0,75H, 1,5H, 2H y 3,5H.

Por otro lado, tomando en cuenta las restricciones 2 y 3, se trató de que el asiento más lejano estuviera a una distancia menor a tres veces y media la altura visible de la pantalla. Dicho asiento fue colocado a aproximadamente 3.4***H***.

Para respetar la restricción 4, se decidió que, en lugar de disponer los asientos en dos filas paralelas al televisor, se colocaran en arcos, y girados de forma tal que la línea de visión de cada observador pasara por el centro de la pantalla. Se ubicaron los asientos de tal forma que, independientemente de la altura de los sujetos, no se obstruyera la línea de visión de los de atrás. Al momento de la sesión los observadores se sentaron de forma aleatoria.

Para los asientos 1, 2, 3, 5 y 6 fue posible respetar las cinco restricciones a la vez. Sin embargo, este no fue el caso para los asientos 4 y 7 (de ángulo γ), para los cuales el ángulo de visión quedó a 40 grados. Cabe destacar que la recomendación ITU-R BT.500-13 [2] dice que el valor del ángulo menor a 30 grados está pensado para televisores de tubos de rayos catódicos y la pantalla utilizada era de cristal líquido.

Finalmente, debió también tenerse en cuenta que la línea de visión de la mayoría de los observadores quedara contenida en un plano ortogonal a la pantalla que la cortara a la mitad de su altura visible. Para esto se colocó el televisor con el comienzo de la sección visible de la pantalla a una altura de 85 centímetros respecto del suelo. En el LATU se colocó en un soporte y en el IIE sobre una mesa.

Los valores finales de distancias de asientos y de ángulos se pueden ver en la Tabla 3. Cabe recordar que la disposición es simétrica respecto de un plano vertical que corta perpendicularmente a la pantalla por el medio, siendo la posición del asiento 1 simétrica a la del 3, la del 4 a la del 7 y la del 5 a la del 6.

Dimensión	Descripción	Valor
D_2	Distancia desde el centro de la pantalla a la posición 2.	1,41m
D_1 , D_3	Distancia desde el centro de la pantalla a la posición 1 o 3.	1,48m
D_5 , D_6	Distancia desde el centro de la pantalla a la posición 5 o 6.	2,28m
D_4 , D_7	Distancia desde el centro de la pantalla a la posición 4 o 7.	2,32m
α	Ángulo entre las posiciones 2 y 6 al centro de la pantalla	14°
β	Ángulo entre las posiciones 2 y 3 al centro de la pantalla	28°
γ	Ángulo entre las posiciones 2 y 7 al centro de la pantalla	40°
θ_2	Giro de asiento 2	0°
θ_1, θ_3	Giro de asientos 1 y 3	23°
θ_5, θ_6	Giro de asientos 5 y 6	10°
θ_4, θ_7	Giro de asientos 4 y 7	37°

Tabla 3: Dimensiones

Como fue dicho previamente, los asientos se colocaron girados, de forma tal que el centro de la pantalla pasara por la línea de visión de los observadores. El ángulo θ_i (con i = 1..7) correspondiente al giro del i-ésimo asiento, se define como el menor ángulo entre la línea formada por sus dos patas traseras y una línea paralela al televisor. Los valores de estos ángulos de giro pueden verse en la Tabla 3.

2.2.5 Secuencias de video originales.

2.2.5.1 Requerimientos

A efectos de diseñar las sesiones de evaluación subjetiva, es necesario seleccionar secuencias de video "originales" llamadas SRC (del inglés *Source Reference Circuit*). Éstas deben cumplir los siguientes requerimientos:

1. No deben tener degradaciones visibles a criterio del grupo experto que define la sesión.
2. El origen de la secuencia debe ser conocido y registrado.
3. Deben tener una duración entre 10 y 12 segundos.
4. Deben cubrir diferentes rangos de complejidad espacial y temporal.

Se evaluaron diecisiete posibles secuencias originales, de las cuales finalmente se seleccionaron catorce, asegurando siempre que se respetaran los requerimientos previamente mencionados. Todas las secuencias consideradas fueron filmadas y/o provistas por la universidad SJTU (Shanghai Jiao Tong University) y expresamente creadas con el propósito de llevar a cabo sesiones de evaluación subjetiva de calidad de video en resolución 4K (3840 x 2160) [11]. Se pueden ver detalles de todas ellas en la Tabla 4. Todas las secuencias tienen una duración de doce segundos y carecen de sonido.

Título	Descripción	Primer Cuadro
Bund Nightscape	Plano aéreo de una ciudad en cámara rápida, durante la noche.	
Campfire Party	Muestra llamas ante el equipo de NERC-DTV en una fiesta con una fogata.	
Construction Field	Muestra una excavadora en un sitio de construcción.	
Fountains	Muestra chorros verticales de una fuente de agua frente a un edificio alto.	
Marathon	Muestra la escena de las primeras etapas de la 2012 Shanghai International Marathon Race.	

Runners	Muestra muchos corredores en medio de la 2012 Shanghai International Marathon Race.	
Rush Hour	Muestra muchos estudiantes yendo a la cantina o al dormitorio después de clases.	
Scarf	Muestra unas bufandas moviéndose al viento.	
Traffic and Building	Muestra tráfico con una ciudad de fondo.	
Traffic Flow	Muestra caminos con automóviles moviéndose en distintas direcciones.	
Tree Shade	Muestra un lugar con mucha vegetación y un árbol cuyas hojas se mueven al viento.	
Wood	Muestra un bosque en el campus de SJTU con rayos del sol penetrándolo.	
Mobile	Muestra un tren de juguete circulando y otros objetos en movimiento.	
Coastguard	Muestra un barco navegando bajo el sol.	

Tabla 4: Secuencias de video utilizadas

2.2.5.2 *Complejidad espacial-temporal*

Según la recomendación ITU-T P.910 [3]: "La ubicación de la secuencia de video dentro de la matriz espacial-temporal es importante ya que la calidad de una secuencia de video transmitida (especialmente después de pasar a través de un códec de baja velocidad binaria) depende a menudo en gran medida de dicha ubicación. […] Por lo general, la dificultad de compresión está directamente relacionada con la información espacial y temporal de una secuencia."

Para cuantificar la complejidad espacial y temporal se utilizaron las siguientes métricas recomendadas en ITU-T P.910 [3]:

- Índice de información espacial (***SI**, del inglés Spatial Information*)
- Índice de información temporal (***TI**, del inglés Temporal Information*)

A continuación, se dan definiciones de estas métricas tal como se las ve en la recomendación ITU-T P.910 [3].

2.2.5.2.1 Índice de Información Espacial (SI)

La información de percepción espacial (**SI**) se basa en el filtro Sobel y se calcula en cuatro pasos:

1. Se filtra cada trama de vídeo (plano de luminancia) en un momento *n* (F_n) con el filtro Sobel [$Sobel(F_n)$].
2. Se calcula la desviación típica de los píxeles ($std_{espacio}$) de cada trama filtrada con el filtro Sobel.
3. Se repite el paso 2 para cada trama de la secuencia de vídeo y esto da por resultado una serie temporal de información espacial de la escena.
4. Se elige el valor máximo de la serie temporal (max_{tiempo}) como representación del contenido de información espacial de la escena.

Este proceso se puede representar en forma de ecuación como sigue:

$$\mathbf{SI} = max_{tiempo}\{std_{espacio}[Sobel(F_n)]\} \qquad (2.1)$$

2.2.5.2.2 Índice de Información Temporal (TI)

La información de percepción temporal (**TI**) se basa en la característica de diferencia de movimiento $M_n(i,j)$, que es la diferencia entre los valores de píxeles (del plano de luminancia) en la misma ubicación en el espacio pero en momentos o tramas sucesivos. $M_n(i,j)$ se define, como una función del tiempo (*n*) de la siguiente manera:

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \qquad (2.2)$$

donde $F_n(i,j)$ es el píxel en la *i*-ésima fila y *j*-ésima columna de la *n*-ésima trama en el tiempo.

La medida de la información temporal (**TI**) se calcula como el valor máximo en el tiempo (max_{tiempo}) de la desviación típica en el espacio ($std_{espacio}$) de $M_n(i,j)$ en todas las *i* y *j*.

$$\mathbf{TI} = max_{tiempo}\{std_{espacio}[M_n(i,j)]\} \quad (2.3)$$

Un mayor movimiento en las tramas adyacentes dará lugar a valores de TI más elevados.

2.2.5.2.3 Selección de SRCs

Se seleccionaron las secuencias de forma tal que se abarcasen distintos grados de complejidad espacial y temporal.

La Figura 2 es una gráfica de ***SI*** contra ***TI*** donde cada punto corresponde a una secuencia del conjunto de candidatas. Los valores de ***SI*** y ***TI*** se obtuvieron con el programa SITI-master [12], y se listan en la Tabla 5.

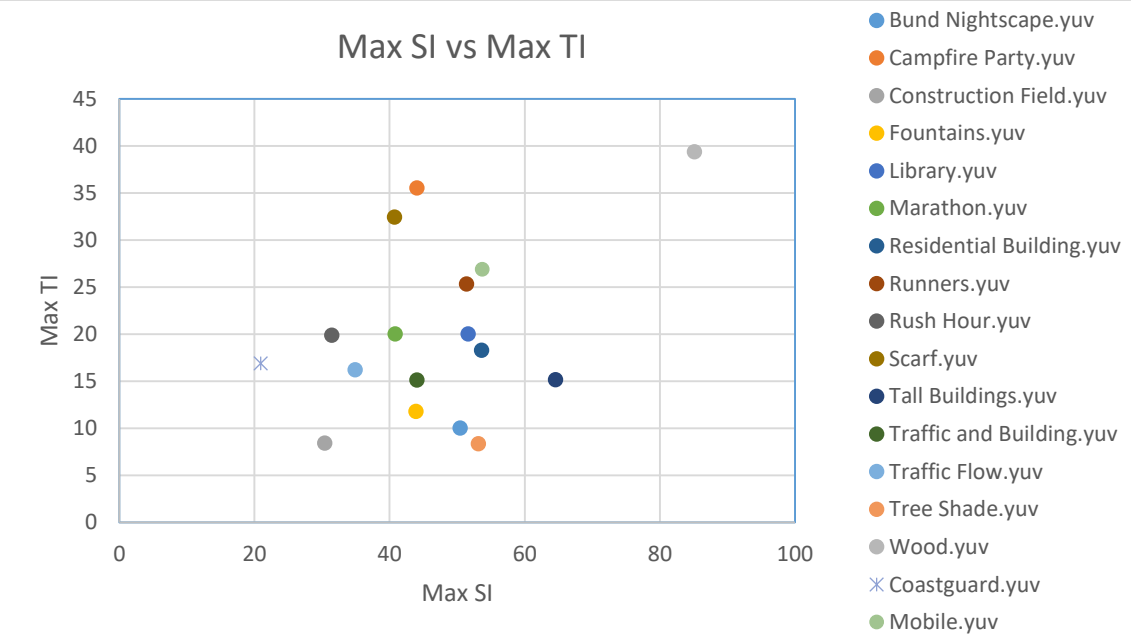


Figura 2: Gráfica SI vs TI

Nombre de la secuencia	Origen	SI Máximo	TI Máximo
Bund Nightscape.yuv	SJTU	50,44	10,06
Campfire Party.yuv	SJTU	44,07	35,55
Construction Field.yuv	SJTU	30,39	8,45
Fountains.yuv	SJTU	43,93	11,82
Library.yuv	SJTU	51,64	20,03
Marathon.yuv	SJTU	40,85	19,05
Residential Building.yuv	SJTU	53,62	18,30
Runners.yuv	SJTU	51,40	25,35
Rush Hour.yuv	SJTU	31,43	19,91
Scarf.yuv	SJTU	40,72	32,44
Tall Buildings.yuv	SJTU	64,59	15,19
Traffic and Building.yuv	SJTU	44,04	15,13
Traffic Flow.yuv	SJTU	34,89	16,22

Tree Shade.yuv	SJTU	53,15	8,38
Wood.yuv	SJTU	85,16	39,38
Coastguard.yuv	Elemental	20,92	16,92
Mobile.yuv	Elemental	53,70	26,91

Tabla 5: Valores de SI y TI para los videos considerados

2.2.6 Circuitos hipotéticos de referencia (HRC)

Según ITU-T REC P.913 [4], se le llama "*Circuito Hipotético de Referencia*" o "HRC" (del inglés *Hypothetical Reference Circuit*) a: "*Una combinación fija de un codificador operando a una tasa de bits dada, una condición de red y un decodificador*."

En el caso de este experimento, no hay una red involucrada, por lo cual esa parte de la definición no aplica. Por otro lado, para todas las compresiones realizadas se utilizó siempre el mismo códec (HEVC). Por tanto, el único factor que distingue a nuestros HRCs entre sí, está relacionado a la tasa de bits.

Al comprimir cada SRC, se buscó obtener secuencias de video procesadas (PVS, del inglés *Processed Video Sequence*) que en nuestra opinión abarcasen un amplio rango de calidades, tal como se hizo en [5] y [6]. A estos efectos, para cada SRC, se seleccionaron manualmente seis tasas de bits distintas para las cuales lográramos diferenciar seis niveles de degradación, uno de los cuales siempre correspondía a una degradación imperceptible. Esto a su vez define seis niveles de tasas de bits, y es en función de ellos que definimos nuestros HRCs. En la Tabla 6 se indica el número de HRC correspondiente a cada uno de estos niveles.

#HRC	Nivel de Tasa de Bits
1	Muy bajo
2	Bajo
3	Medio bajo
4	Medio alto
5	Alto
6	Muy alto

Tabla 6: HRCs

2.2.7 Secuencias de video procesadas (PVS)

Para la realización de las evaluaciones subjetivas se generaron secuencias de video degradadas, partiendo de los videos originales (SRC) y procesándolos según los distintos circuitos de referencia (HRC) establecidos. Las secuencias de video generadas se denominan PVS.

A estos efectos, se utiliza el programa FFmpeg [13], al cual se le dio como entrada un video con formato raw (YUV), y el cual dio como salida un archivo mp4 codificado en HEVC según el circuito de referencia utilizado. Dicho programa se ejecuta por línea de comandos y los comandos utilizados tienen la siguiente forma:

```
ffmpeg.exe -f rawvideo -video_size 3840x2160 -pixel_format yuv420p -i <entrada>.yuv -c:v hevc -b:v <bitRate>k -maxrate <bitRate>k -minrate <bitRate>k -r 30 -x265-params -an -y <salida>.mp4
```

- `-f rawvideo -video_size 3840x2160 -pixel_format yuv420p` – Especifica el formato del video de entrada, necesario porque la entrada es un video crudo.
- `-i <entrada>.yuv` – Especifica el archivo de entrada.
- `-c:v hevc` – Especifica el códec de video, en este caso HEVC.
- `-b:v <bitRate>k -maxrate <bitRate>k -minrate <bitRate>k` – Especifica la tasa de bits de codificación.
- `-r 30` – Especifica la tasa de cuadros por segundo.
- `-x265-params` – Generalmente, se pasan opciones a x265 con el argumento "-x265 -params". Para afinar el proceso de codificación se puede por tanto pasar cualquier opción que esté listada en la documentación de x265 [14].
- `-an` – Salida sin sonido.
- `-y <salida>.mp4` – Especifica el archivo de salida.

Para cada entrada, se utilizaron los valores de tasa de bits de la Tabla 7 obteniéndose seis versiones con distintos grados de degradación.

SRC	#SRC	#HRC	#PVS	PVS	Tasa de bits (kbps)
Bund Nightscape.yuv	1	1	1	Bund_Nightscape1.mp4	2341,39
		2	2	Bund_Nightscape2.mp4	5069,66
		3	3	Bund_Nightscape3.mp4	7964,73
		4	4	Bund_Nightscape4.mp4	11994,09
		5	5	Bund_Nightscape5.mp4	18011,08
		6	6	Bund_Nightscape6.mp4	24072,74
Campfire Party.yuv	2	1	7	Campfire_Party1.mp4	1908
		2	8	Campfire_Party2.mp4	3566
		3	9	Campfire_Party3.mp4	5916
		4	10	Campfire_Party4.mp4	7817
		5	11	Campfire_Party5.mp4	12300,38
		6	12	Campfire_Party6.mp4	16231,04
Coastguard.yuv	3	1	13	Coastguard1.mp4	1491,21
		2	14	Coastguard2.mp4	1996,5
		3	15	Coastguard3.mp4	3498,1
		4	16	Coastguard4.mp4	5956,84
		5	17	Coastguard5.mp4	9886,47
		6	18	Coastguard6.mp4	11882,27
Construction Field.yuv	4	1	19	Construction_Field2.mp4	927,59
		2	20	Construction_Field3.mp4	1432,58
		3	21	Construction_Field4.mp4	2420,89

		4	22	Construction_Field5.mp4	3917,54
		5	23	Construction_Field6.mp4	9869,63
		6	24	Construction_Field7.mp4	14877,93
Fountains.yuv	5	1	25	Fountains1.mp4	2404,09
		2	26	Fountains2.mp4	5240,66
		3	27	Fountains3.mp4	8167,25
		4	28	Fountains4.mp4	12190,31
		5	29	Fountains5.mp4	18221,92
		6	30	Fountains6.mp4	24236,71
Marathon.yuv	6	1	31	Marathon1.mp4	2054,05
		2	32	Marathon2.mp4	4056,19
		3	33	Marathon3.mp4	7514,98
		4	34	Marathon4.mp4	11057,18
		5	35	Marathon5.mp4	15046,79
		6	36	Marathon6.mp4	18043,88
Mobile.yuv	7	1	37	Mobile1.mp4	821,53
		2	38	Mobile2.mp4	1111,46
		3	39	Mobile3.mp4	1788,37
		4	40	Mobile4.mp4	2323,05
		5	41	Mobile5.mp4	5377,13
		6	42	Mobile6.mp4	8447,09
Runners.yuv	8	1	43	Runners1.mp4	2124,51
		2	44	Runners2.mp4	5354,24
		3	45	Runners3.mp4	6753,5
		4	46	Runners4.mp4	9930,42
		5	47	Runners5.mp4	12320,76
		6	48	Runners6.mp4	18263,62
Rush Hour.yuv	9	1	49	Rush_Hour1.mp4	1573,02
		2	50	Rush_Hour2.mp4	2737,13
		3	51	Rush_Hour3.mp4	4798,39
		4	52	Rush_Hour4.mp4	5988,94
		5	53	Rush_Hour5.mp4	8126,02
		6	54	Rush_Hour6.mp4	12287,15
Scarf.yuv	10	1	55	Scarf0.mp4	377,06
		2	56	Scarf1.mp4	1537,41
		3	57	Scarf2.mp4	2886,03
		4	58	Scarf3.mp4	5111,91
		5	59	Scarf4.mp4	7025,3
		6	60	Scarf5.mp4	11723,19
Traffic And Building.yuv	11	1	61	TrafficAndBuilding0.mp4	1087,67
		2	62	TrafficAndBuilding1.mp4	2256,33
		3	63	TrafficAndBuilding2.mp4	4966,11

		4	64	TrafficAndBuilding3.mp4	7845,67
		5	65	TrafficAndBuilding4.mp4	11828,24
		6	66	TrafficAndBuilding5.mp4	17838,98
Traffic Flow.yuv	12	1	67	Traffic_Flow1.mp4	899,09
		2	68	Traffic_Flow2.mp4	1959,23
		3	69	Traffic_Flow3.mp4	4009,65
		4	70	Traffic_Flow4.mp4	5000,24
		5	71	Traffic_Flow5.mp4	7975,35
		6	72	Traffic_Flow6.mp4	11850,27
Tree Shade.yuv	13	1	73	Tree_Shade0.mp4	1019,83
		2	74	Tree_Shade1.mp4	2218,25
		3	75	Tree_Shade2.mp4	4871,96
		4	76	Tree_Shade3.mp4	7627,09
		5	77	Tree_Shade4.mp4	11475,88
		6	78	Tree_Shade5.mp4	17325,7
Wood.yuv	14	1	79	Wood1.mp4	1779,16
		2	80	Wood2.mp4	3689,95
		3	81	Wood3.mp4	5180,78
		4	82	Wood4.mp4	8090,87
		5	83	Wood5.mp4	11952,7
		6	84	Wood6.mp4	14928,66

Tabla 7: Secuencias de video procesadas

2.3 Método de Evaluación

El método a utilizar en las pruebas es el denominado ACR (del inglés *Absolute Category Rating*), también conocido como "Índice por categoría absoluta". En esta sección se presenta una descripción de dicho método, extraída de la recomendación ITU-T P.910 [3].

También se presenta una breve descripción de la aplicación usada que implementa este método.

2.3.1 Descripción del método ACR

El método de los índices por categorías absolutas (ACR) es un juicio de categorías en el que las secuencias de prueba se presentan una por vez y se califican independientemente en una escala de categorías. (Este método se denomina también método de evaluación con un solo estímulo.) El método especifica que después de cada presentación se invite a los sujetos a evaluar la calidad de la secuencia mostrada. En la Figura 3 se ilustra el diagrama de tiempos de la presentación del estímulo. Si se utiliza un tiempo de votación constante entonces el tiempo de votación debe ser igual o inferior a 10 s.

En la Figura 3 (extraída de ITU-T P.910 [3]) se ilustra el diagrama de tiempos de la presentación del estímulo.

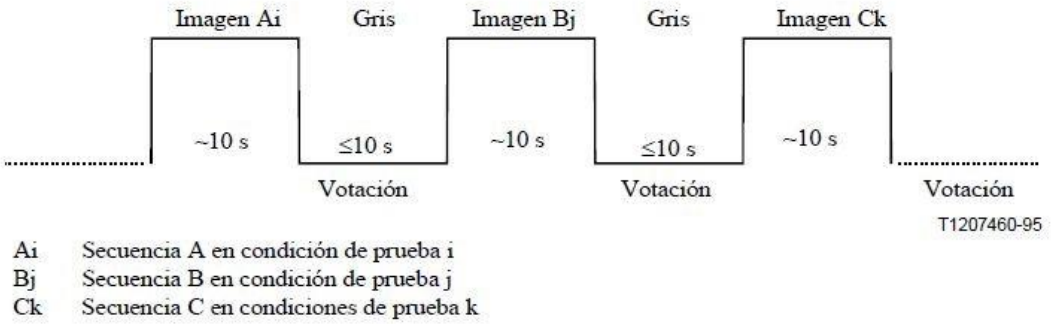


Figura 3: Presentación del estímulo en el método ACR

El tiempo de votación deberá ser igual o inferior a 10 s. El tiempo de presentación puede reducirse o aumentarse en función del contenido del material de prueba. Para evaluar la calidad global se debe utilizar la siguiente escala de cinco niveles, como se muestra en la Tabla 8:

Puntuación	Nivel
5	Excelente
4	Buena
3	Aceptable
2	Mediocre
1	Mala

Tabla 8: Puntajes y Niveles

2.3.2 Descripción de la aplicación utilizada para evaluar

A efectos de realizar las sesiones de evaluación, fue necesario disponer de una aplicación que tuviera las siguientes funciones:

- Armar una lista de reproducción aleatoria en base a las PVSs seleccionadas para la sesión. Se hace especial hincapié en que la reproducción debe ser aleatoria de manera que los efectos del cansancio o la adaptación sobre las evaluaciones se equilibren de una sesión a otra [2].
- Llamar a un programa que reproduzca la lista de PVSs.
- Solicitar al observador su calificación en una escala de cinco valores (malo, mediocre, aceptable, bueno, excelente) tras la reproducción de cada PVSs.
- Aguardar a recibir las calificaciones de todos los observadores para una PVS antes de empezar a reproducir la siguiente.
- Permitir determinar manualmente la cantidad de participantes de la sesión. Esto fue útil en sesiones en las que faltaron observadores.
- Permitir a los participantes conectarse desde dispositivos móviles para ingresar sus calificaciones.

2.3.2.1 Software utilizado

2.3.2.1.1 Descripción

Para reproducir los videos y recolectar las votaciones de los participantes, se utilizó una aplicación proporcionada por nuestros tutores. Esta aplicación fue desarrollada en el marco del proyecto VQI ([15], [16] y [10]). La misma consta de tres módulos que se detallan a continuación y cuyo esquema se presenta en la Figura 4.

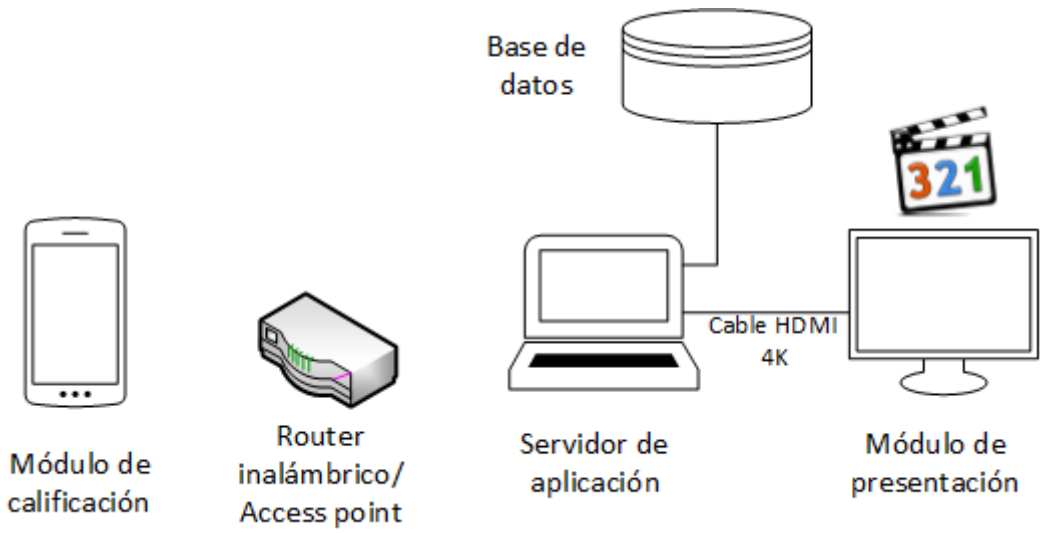


Figura 4: Esquema de conexión de la aplicación

Servidor de Aplicación: Contiene la lógica para administrar las sesiones de pruebas subjetivas, controla la ejecución aleatoria de videos, y almacena la información obtenida de las pruebas en una base de datos. En la Figura 5 se muestra la interfaz para la gestión de las sesiones.

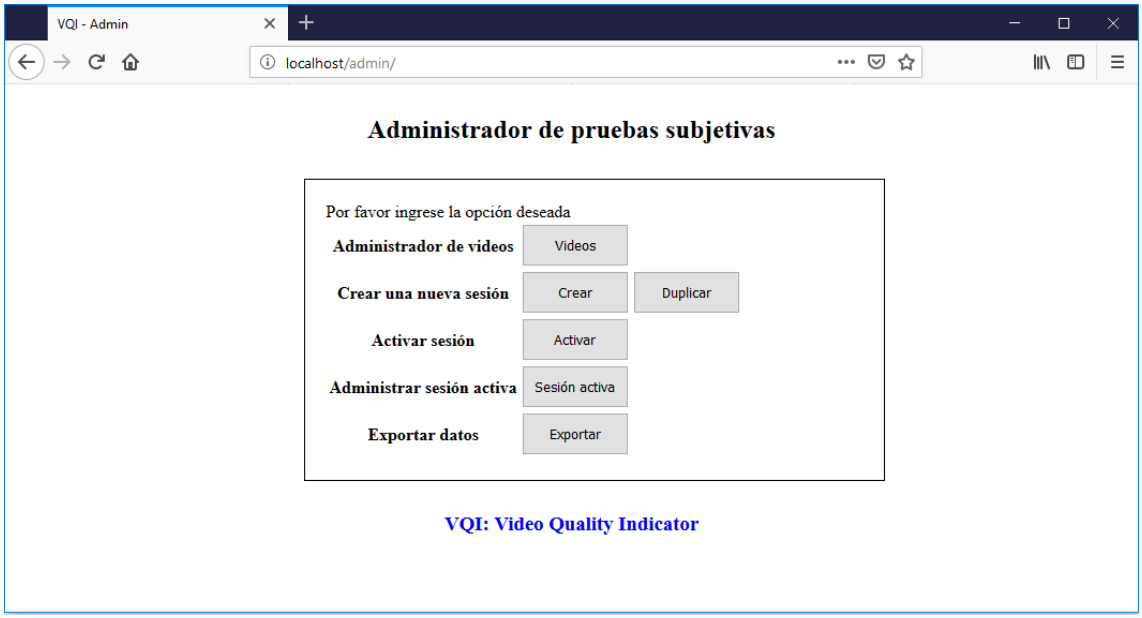


Figura 5: Administración de sesiones de prueba

Debajo se detallan las distintas opciones que permite este módulo.

- "Administrador de videos": permite altas y bajas de los videos a utilizar en una sesión.
- "Crear una nueva sesión": permite crear una nueva sesión eligiendo los videos a mostrar o duplicar una sesión existente.
- "Activar sesión": permite activar una sesión creada y definir la cantidad de participantes. Además, permite iniciar y pausar la reproducción de los videos.
- "Administrar sesión activa": permite activar o desactivar una sesión de prueba.
- "Exportar datos": permite exportar las votaciones y los datos de los participantes.

Módulo de Presentación: este módulo es el encargado de presentar los videos a ser evaluados. Permite configurar un reproductor que se va a utilizar, que en nuestro caso fue Media Player Classic, versión 1.7.13 (e37826845).

Módulo de Calificación: Consiste en un servicio web optimizado para dispositivos móviles, que permite a los evaluadores identificarse y enviar las calificaciones de los videos vistos al servidor.

Al comienzo de la prueba, este módulo solicita los datos al observador que fueron detallados en la sección 2.2.1. Si el observador ya está registrado, por haber realizado una prueba anteriormente, solo solicita el número de cédula.

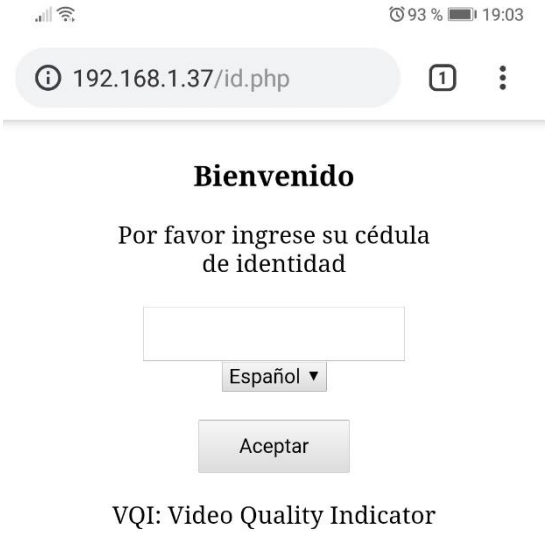


Figura 6: Ingreso de datos por parte de los observadores

Durante la sesión de prueba, solicita a los observadores la calificación del video que acaban de ver, como se ve en la Figura 7.

Figura 7: Ingreso de calificación de los observadores

2.3.2.1.2 Funcionamiento

Antes de comenzar la sesión, a través de la interfaz web de administración de la aplicación, se crea una nueva sesión (o se duplica una existente), definiendo los videos (PVSs) que van a mostrarse durante la prueba. Luego, se procede a activar la sesión y a definir la cantidad de participantes.

Los participantes se conectan utilizando móviles de prueba. Al comienzo se les piden los datos, y una vez que todos los participantes completaron los datos requeridos, se puede comenzar la prueba.

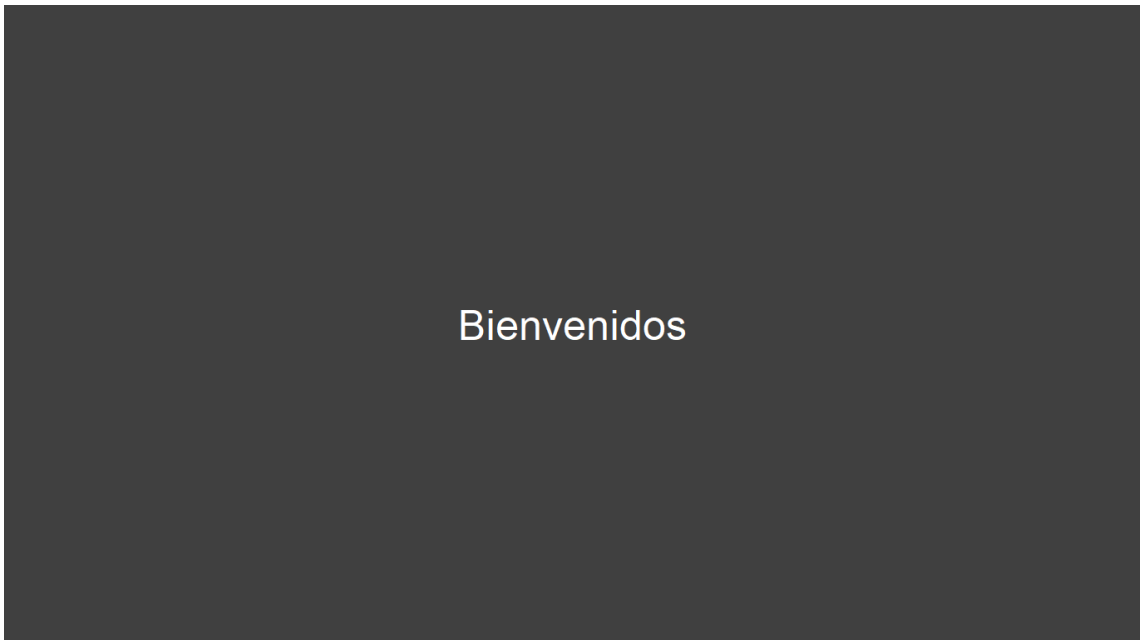


Figura 8: Pantalla de bienvenida

Inicialmente, en el televisor se muestra un cartel de bienvenida como el que se ve en la Figura 8. Cuando comienza la prueba y durante todo el transcurso de la misma, se indica el comienzo de un nuevo video y cuando el mismo termina, se solicita a los observadores que voten el video que acaban de ver. En ese momento, aparece en sus celulares la escala para calificar el video, vista en la Figura 7. El siguiente video no se reproduce hasta que todos los participantes hayan emitido su voto.

Finalmente, cuando todos los videos fueron reproducidos y calificados, se indica el final de la prueba en el televisor agradeciendo a los participantes.

2.3.2.1.3 Cambios realizados

En nuestro proyecto fue necesario modificar la aplicación para que pida el número de asiento a cada participante, lo cual no había sido necesario en proyectos anteriores. Además, si el participante ya está registrado, no solo se pide el número de cédula, sino que también se pide el número de asiento nuevamente. Aunque esta última funcionalidad no fue necesaria, ya que cada observador tomó parte en una sola sesión útil, se consideró que podía ser de utilidad en proyectos futuros donde un mismo observador tomara parte en más de una sesión.

Los cambios realizados se detallan a continuación:

1. Base de datos:
    - Se agrega la columna *position* en la tabla *participants.*

2. Se agrega el paso que pide el asiento al participante:
    - C:\xamp\htdocs\position.php
3. Se modifican los pasos ya existentes para tomar en cuenta el pedido del asiento en los siguientes archivos:
    - C:\xampp\htdocs\id.php
    - C:\xampp\htdocs\userdata1.php
    - C:\xampp\htdocs\userdata2.php
    - C:\xampp\htdocs\userdata3.php
    - C:\xampp\htdocs\userdata4.php
4. Se agregan los textos necesarios en inglés y español para pedir el asiento al participante en el archivo:
    - C:\xampp\htdocs\configuration.ini
5. Se modifican los archivos donde se guardan los datos de los participantes para que también se guarde la posición:
    - C:\xampp\htdocs\server\existUser.php
    - C:\xampp\htdocs\server\insertUser.php
6. Se modifica el archivo de exportación de los datos para que incluya la posición de cada participante:
    - C:\xampp\htdocs\export\generator.php

El archivo nuevo necesario y los que fueron modificados se adjuntan en el CD que acompaña esta documentación.

2.3.2.2 Hardware utilizado

Para correr esta aplicación, se utilizó la computadora conectada por medio de un cable HDMI 4K al televisor (descriptos en 2.2.3.1 y 2.2.3.2). La aplicación requería que tanto los participantes como la computadora estuvieran conectados a una misma red wifi para que los dispositivos móviles utilizados vieran la computadora. De manera de evitar que llegaran mensajes o notificaciones a los dispositivos utilizados por los participantes, se configuró un *access point* sin acceso a internet, utilizando el router detallado en 2.2.3.3.

2.4 Análisis de los resultados

2.4.1 Cálculo de MOS

Dada una PVS *j*, se define MOS_j (del inglés *Mean Opinion Score*) como el promedio de las puntuaciones que todos los observadores le asignan. El primer paso en el análisis de los resultados es calcular estos promedios, dados por:

$$MOS_j \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} u_{ij} \quad (2.4)$$

Donde:

- u_{ij} representa la puntuación del observador *i* para la PVS *j*
- *N* es la cantidad de observadores.

En nuestro caso, se calculan valores de MOS promediando las opiniones de distintos observadores ubicados en la misma posición. En principio, no se calcula el MOS como promedio de opiniones de observadores en distintas posiciones.

2.4.2 Validaciones de calificaciones individuales

A efectos de determinar si tiene sentido considerar las puntuaciones individuales asignadas por un observador, es necesario determinar cuánto éstas se correlacionan con las puntuaciones promedio de todos los evaluadores de la misma posición. Para analizar dicha correlación, la Unión Internacional de Telecomunicaciones recomienda en ITU-T REC P.913 [4] utilizar el coeficiente de correlación lineal de Pearson.

El coeficiente de correlación lineal de Pearson entre dos vectores de datos ***x*** e ***y*** de largo ***K***, se define como:

$$LPCC(\boldsymbol{x},\boldsymbol{y}) \stackrel{\text{def}}{=} \frac{\sum_{j=1}^{K} x_j y_j - \frac{\left(\sum_{j=1}^{K} x_j\right)\left(\sum_{j=1}^{K} y_j\right)}{K}}{\sqrt{\left(\sum_{j=1}^{K} {x_j}^2 - \frac{\left(\sum_{j=1}^{K} x_j\right)^2}{K}\right)\left(\sum_{j=1}^{K} {y_j}^2 - \frac{\left(\sum_{j=1}^{K} y_j\right)^2}{K}\right)}} \quad (2.5)$$

Dicho coeficiente puede tomar valores entre -1 y 1. Un valor absoluto de 1 quiere decir que una ecuación lineal describe la relación entre ***x*** e ***y*** a la perfección. Un valor de 0 implica que no hay correlación lineal entre los vectores de datos.

Para descartar las opiniones de un observador *i*, se verifica que se cumplan dos criterios de correlación. Uno de ellos toma en cuenta las opiniones de un observador de cada PVS y los MOS de cada PVS, mientras que el otro criterio analiza los promedios de las opiniones del observador en cada HRC y los promedios de los MOS de todas las PVSs para ese HRC.

2.4.2.1 Correlación según PVSs

Para analizar esta correlación, dado un observador *i* y un número total $\boldsymbol{P}$ de PVSs, tomamos vectores $\boldsymbol{x}_1$ e $\boldsymbol{y}_1$ de dimensión $\boldsymbol{K}_1$ tales que:

$$\boldsymbol{x_1}[j] = MOS_j \quad (2.6)$$

$$\boldsymbol{y_{1i}}[j] = u_{ij} \quad (2.7)$$

$$\boldsymbol{K_1} = \boldsymbol{P} \quad (2.8)$$

Donde recordamos que:

- *j* indica el número de PVS
- u_{ij} es la puntuación que el *i*-ésimo observador asigna a la *j-ésima PVS*

Y calculamos el coeficiente de correlación de Pearson entre estos dos vectores, al cual denominamos $\boldsymbol{r_1}(\mathrm{i})$, esto es:

$$\boldsymbol{r_1}(\mathrm{i}) = \mathrm{LPCC}(\boldsymbol{x_1}, \mathbf{y_{1i}}) \quad (2.9)$$

2.4.2.2 Correlación según HRCs

Previo a este análisis, es útil definir a priori dos índices *m* y *n*, tales que *m* indica el número de SRC, y *n* indica el número de HRC, involucrados en la definición de una PVS dada. Estos índices proveen una alternativa al índice *j* (utilizado en la parte anterior) que indicaba el número de PVS. En base a esto, podemos utilizar MOS_{mn} para indicar el MOS de la PVS definida por la SRC *m* y el HRC *n* y u_{imn} para indicar la puntuación que del *i*-ésimo observador asigna a dicha PVS.

Sea L el número total de SRCs. Definimos el concepto de "MOS de todos los observadores condicionado a un HRC *n*", denotado por $CMOS(n)$ y dado por:

$$CMOS(n) \stackrel{\text{def}}{=} \frac{1}{L}\sum_{m=1}^{L} MOS_{mn} \quad (2.10)$$

Definimos también el concepto de "MOS de un observador *i* condicionado a un HRC *n*", esto es, el promedio de todas las puntuaciones que un observador *i* asigna a PVSs creadas con un mismo HRC *n*, denotado por $CMOS_i(n)$ y dado por:

$$CMOS_i(n) \stackrel{\text{def}}{=} \frac{1}{L}\sum_{m=1}^{L} u_{imn} \quad (2.11)$$

Para obtener la correlación según HRCs, dado un observador *i* y un número total $\boldsymbol{H}$ de HRCs, tomamos vectores $\boldsymbol{x}_2$ e $\boldsymbol{y}_2$ de dimensión $\boldsymbol{K}_2$ tales que:

$$\boldsymbol{x}_2[n] = CMOS(n) \quad (2.12)$$

$$\boldsymbol{y}_{2i}[n] = CMOS_i(n) \quad (2.13)$$

$$\boldsymbol{K}_2 = \boldsymbol{H} \quad (2.14)$$

Y calculamos el coeficiente de correlación de Pearson entre estos dos vectores, al cual denominamos $\boldsymbol{r_2}(\mathrm{i})$, esto es:

$$\boldsymbol{r_2}(\mathrm{i}) = \mathrm{LPCC}(\boldsymbol{x_2}, \mathbf{y_{2i}}) \quad (15)$$

La razón para analizar la correlación según HRCs haciendo uso del coeficiente de correlación $\boldsymbol{r_2}(\mathrm{i})$ es que un observador puede tener una preferencia individual de contenido que difiere de las de otros observadores. Tal preferencia causaría que $\boldsymbol{r_1}(\mathrm{i})$ decreciera, aunque este sujeto podría haber votado consistentemente. El análisis por HRC promedia la preferencia de contenido de un individuo y chequea consistencia a través de distintas condiciones de error. Un ejemplo de esto se da en la sección 3.3.3.

2.4.2.3 Criterio utilizado para descartar observadores

Se utiliza el criterio para descartar observadores recomendado en ITU-T REC P.913 [4] para sesiones de evaluación subjetivas que usan el método ACR. Este criterio establece que ha de descartarse a todo observador *i* para el cual se cumplan simultáneamente las siguientes dos condiciones:

$$\boldsymbol{r_1}(i) < 0{,}75 \quad (2.16)$$

$$\boldsymbol{r_2}(i) < 0{,}8 \quad (2.17)$$

Tales observadores deben ser descartados de a uno, empezando por aquel con el mayor promedio de las cantidades por las cuales los umbrales exceden a sus respectivos coeficientes, y luego volviendo a calcular $\boldsymbol{r_1}(\mathrm{i})$ y $\boldsymbol{r_2}(\mathrm{i})$ para todos los observadores restantes.

2.4.3 Cálculo de intervalos de confianza

Cuando se presentan los resultados de una sesión de evaluación, cada MOS_j debe tener un intervalo de confianza asociado, que se obtiene a partir de la desviación típica y la cantidad de observadores. Asumiendo una distribución normal para el conjunto de las puntuaciones de cada observador para una PVS *j* dada, se utiliza un intervalo de confianza del 95%, que viene dado por las ecuaciones siguientes:

$$IC = \left[MOS_j - \delta_j, MOS_j + \delta_j\right] \quad (2.18)$$

con

$$\delta_j = 1{,}96\frac{S_j}{\sqrt{N}} \quad (2.19)$$

y

$$S_j = \sqrt{\sum_{i=1}^{N} \frac{(MOS_j - u_{ij})^2}{N-1}} \quad (2.20)$$

Donde:

- *i* es el índice del observador y *j* es el índice de la PVS
- u_{ij} representa la puntuación del observador *i* para la PVS *j*
- *N* es la cantidad de observadores.

Con una probabilidad del 95%, el valor absoluto de la diferencia entre el MOS_j experimental y el MOS_j "verdadero" (para un número de observadores muy elevado) es menor que el intervalo de confianza del 95%.

Calculando los intervalos de confianza para cada PVS *j* en cada asiento, se ve cuán similar es la percepción de calidad de un observador que se encuentra en una posición, a la de un observador que se encuentra en otra.

2.4.4 Test t de Student

En [1] se presenta un análisis más riguroso que el presentado en la sección anterior, que consiste en realizar un test t de Student de dos muestras y varianzas desiguales, usando una distribución de dos colas. En nuestro caso, este test se usa para determinar si las calidades subjetivas dadas por los valores promedio de las muestras de un par de asientos no son iguales. La hipótesis nula H_0 en este caso sería que los observadores en diferentes asientos perciben la misma calidad para una PVS dada, y la hipótesis alternativa H_a es que los observadores en distintos asientos no perciben la misma calidad para dicha PVS. Para comparar las medias de dos poblaciones, puede usarse la estadística t, la cual se expresa como:

$$t = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{{s_1^2}/{n_1} + {s_2^2}/{n_2}}} \quad (2.21)$$

donde $\overline{X}_i, s_i^2, n_i$ denotan la media, la varianza y el tamaño de la i-ésima muestra siendo *i = 1,2*.

El cálculo de dichos valores, se hace utilizando las siguientes fórmulas:

$$\overline{X}_i = \frac{1}{n_i} \sum_{k=1}^{k=n_i} X_k \quad (2.22)$$

$$s_i^2 = \frac{1}{n-1} \sum_{k=1}^{k=n_i} (X_k - \overline{X}_i)^2 \quad (2.23)$$

donde X_k es la k-ésima opinión en un asiento para una PVS e *i = 1,2*.

Al calcular la estadística t de esta forma y aproximándola con una distribución t de Student cuyo grado de libertad *DF* viene dado por:

$$DF = \frac{\left({s_1^2}/{n_1} + {s_2^2}/{n_2}\right)^2}{\frac{\left({s_1^2}/{n_1}\right)^2}{n_1 - 1} + \frac{\left({s_2^2}/{n_2}\right)^2}{n_2 - 1}} \quad (2.25)$$

se puede calcular un valor de probabilidad *p* a partir de la estadística t, que indica el grado al cual las medias de las dos poblaciones se consideran diferentes. Cuanto más pequeño es el valor *p*, más significativa es la diferencia entre las distribuciones de las poblaciones.

Un valor *p* menor a 0,05 indica una probabilidad muy baja de cometer un error tipo I (esto es, rechazar la hipótesis nula cuando es cierta). En tal caso, la hipótesis nula puede ser rechazada con confianza, y puede concluirse que hay significado estadístico en que las dos ubicaciones perciben calidades diferentes. Un valor *p* mayor o igual a 0,05 significa que la hipótesis nula no puede ser rechazada con confianza. Sin embargo, todavía existe la posibilidad de cometer un error tipo II (esto es, no poder rechazar la hipótesis nula cuando de hecho la hipótesis alternativa es cierta).

El test será aplicado a todas las PVSs, tomando pares de posiciones. Cada posición será comparada con la posición central, que es la 2 en la Figura 1. Con este criterio se obtendrán 504 resultados del test.

3 Capítulo 3: Sesiones de Evaluación

Se realizaron un total de 16 sesiones, donde participaron 109 observadores. La primera sesión se realizó el 25 de octubre de 2018 en la Sala 6 del edificio Los Talas del LATU, mientras que las 15 sesiones restantes fueron realizadas entre el 22 de noviembre de 2018 y el 14 de diciembre de 2018 en la Sala de Seminarios del Instituto de Ingeniería Eléctrica de la Facultad de Ingeniería (UDELAR). Las fechas, horas y cantidad de participantes de cada prueba se detallan en la Tabla 9.

Fecha	Hora	Lugar	Cantidad de participantes
25/10/2018	19:00	LATU	7
22/11/2018	18:00	IIE - FING	7
26/11/2018	18:00	IIE - FING	7
28/11/2018	16:30	IIE - FING	7
28/11/2018	18:00	IIE - FING	7
29/11/2018	18:00	IIE - FING	7
30/11/2018	18:00	IIE - FING	7
03/12/2018	18:00	IIE - FING	7
04/12/2018	18:00	IIE - FING	7
05/12/2018	18:00	IIE - FING	7
06/12/2018	18:00	IIE - FING	7
07/12/2018	18:00	IIE - FING	7
11/12/2018	16:30	IIE - FING	7
12/12/2018	18:00	IIE - FING	7
13/12/2018	18:00	IIE - FING	5
14/12/2018	18:00	IIE - FING	6

Tabla 9: Sesiones de prueba

Cada sesión de evaluación tuvo una duración de aproximadamente una hora. En los primeros 10 a 15 minutos se realizaba a cada participante un test de visión y uno de daltonismo. Luego, se les daban instrucciones sobre la prueba y la aplicación a utilizar. Finalmente, se procedía a la realización de la prueba, que consistía en ver y calificar 84 videos sin sonido de 12 segundos cada uno. Esta última parte tenía una duración de aproximadamente 40 minutos.

Todos los observadores participaron voluntariamente de las pruebas. La forma de reclutamiento de los mismos fue por medio de redes de amigos, vecinos y conocidos de cada participante del proyecto. Al finalizar cada sesión de prueba se le daba a cada participante una galletita bañada de chocolate.

En las secciones siguientes se muestran los resultados de las pruebas de visión y daltonismo, y se comentan los datos demográficos de los participantes.

Figura 9: Primera sesión de pruebas subjetivas en la sala 6 del edificio Los Talas del LATU

Figura 10: Quinta sesión de pruebas subjetivas en la sala de seminarios del IIE - FING

3.1 Test de visión y daltonismo

Como se comentó anteriormente, antes de llevar a cabo cada sesión de evaluación subjetiva de calidad de video, a cada participante se le realizaba un test de visión y otro de daltonismo como se describió en la sección 2.2.1. En la Tabla 10, se presenta la distribución por asiento de observadores con deficiencia de agudeza visual o daltonismo.

Asiento	Total de sujetos	Visión inferior a 20/30	Daltónicos
1	15	1	2
2	16	0	1
3	16	1	0
4	15	0	0
5	16	0	1
6	16	0	1
7	15	1	2

Tabla 10: Resultados de las pruebas de visión y daltonismo

En ITU-T REC P.913 [4] se recomienda descartar las puntuaciones de aquellos observadores que presentan problemas de visión. Esto es, que su visión esté por debajo de 20/30 o que cometa un error en más de una placa de daltonismo. Sin embargo, nosotros solo descartamos a estos sujetos en los casos donde resultaron ser sujetos atípicos (outliers), como se hizo en [17], donde se analiza la influencia de los sujetos y el entorno en pruebas audiovisuales subjetivas.

3.2 Datos de los observadores

En cada sesión de evaluación, la aplicación usada (descripta en 2.3.2) solicitaba a los participantes los datos presentados en la sección 2.2.1. Las respuestas de los mismos por asiento se detallan en las siguientes tablas.

		Asiento							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Sexo	**Femenino**	6	7	6	5	8	5	3	40
	Masculino	9	9	10	10	8	11	12	69

Tabla 11: Cantidad de observadores según su sexo

En nuestro caso, el 63,3% de los participantes fueron hombres y el 36,7% restante fueron mujeres. Esto difiere de lo que se recomienda en ITU-T REC P.913 [4], que es que aproximadamente el 50% de los sujetos sean hombres y el 50% sean mujeres, si no se especifica lo contrario durante el diseño de la prueba.

		Asiento							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Nivel educativo	**Primaria**	0	0	0	0	1	0	0	1
	Secundaria	3	8	6	2	2	5	4	30
	Terciaria	12	8	10	13	13	11	11	78

Tabla 12: Cantidad de observadores según su nivel educativo

En cuanto al nivel educativo de los participantes:

- el 71,6% tiene nivel terciario
- el 27,5% tiene nivel de secundaria
- el restante 0,9% nivel de primaria

		Asiento						
		1	**2**	**3**	**4**	**5**	**6**	**7**
Edad	**Mínima**	15	17	20	25	21	23	19
	Máxima	57	70	75	63	75	59	58
	Promedio	34,47	39,19	44,44	37,13	36,31	34,88	36,13

Tabla 13: Edades de los observadores

La edad mínima de los participantes fue de 15 años y la máxima de 75 años. El promedio general de edad (en el conjunto de todos los asientos) fue 38 años. En la Tabla 14 se muestra un desglose por franja etaria de los 109 participantes de las pruebas subjetivas realizadas.

Franja	**Cantidad de sujetos**
Edad hasta 20 años	5
Edad entre 21 y 30 años	37
Edad entre 31 y 40 años	35
Edad entre 41 y 50 años	13
Edad entre 51 y 60 años	10
Edad mayor a 61 años	9

Tabla 14: Franjas etarias de los observadores

Con respecto a las edades de los observadores, en ITU-T REC P.913 [4] se recomienda una "distribución de edades bien balanceada". Lo que puede observarse en la Tabla 14 es que en nuestro caso cubrimos un amplio rango de edades que va desde 15 a 75 años.

Por último, se preguntó a los observadores la cantidad de horas que ven contenido audiovisual en diferentes dispositivos. En las siguientes tablas (Tabla 15, Tabla 16, Tabla 17 y Tabla 18) se puede ver la cantidad de observadores por tipo de dispositivo y tiempo invertido. También se muestran gráficamente los resultados por tipo de dispositivo sobre el total de observadores, esto es sobre los 109 participantes.

		Asiento							**Total**	**Porcentaje**
		1	**2**	**3**	**4**	**5**	**6**	**7**		
Teléfonos celulares	**Menos de 1 hora**	4	4	5	4	3	3	5	28	26%
	Entre 1 y 2 horas	4	4	2	4	6	5	4	29	27%
	Más de 2 horas	5	5	5	3	4	6	4	32	29%

	No utilizan	2	3	4	4	3	2	2	20	18%

Tabla 15: Cantidad de observadores según el tiempo de visualización de contenido audiovisual en teléfono celular

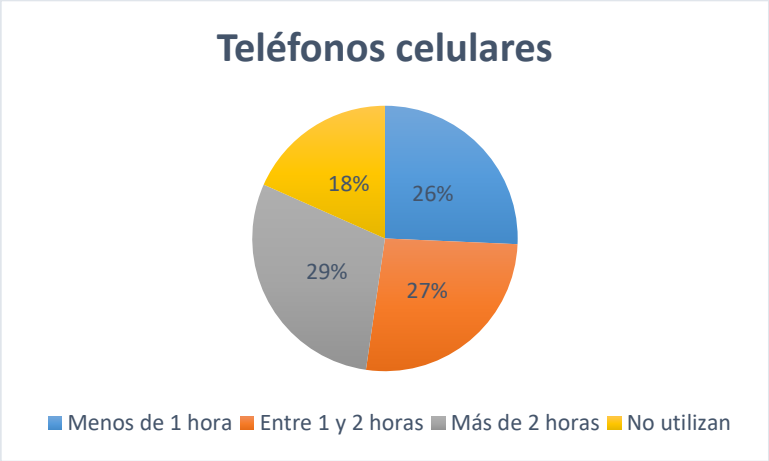


Figura 11: Porcentajes de sujetos según el tiempo de utilización de teléfonos celulares

		Asiento							Total	Porcentaje
		1	**2**	**3**	**4**	**5**	**6**	**7**		
Tablets	**Menos de 1 hora**	4	3	5	1	2	5	2	22	20%
	Entre 1 y 2 horas	0	1	0	0	0	0	1	2	2%
	Más de 2 horas	0	0	0	0	0	0	0	0	0%
	No utilizan	11	12	11	14	14	11	12	85	78%

Tabla 16:: Cantidad de observadores según el tiempo de visualización de contenido audiovisual en tablets

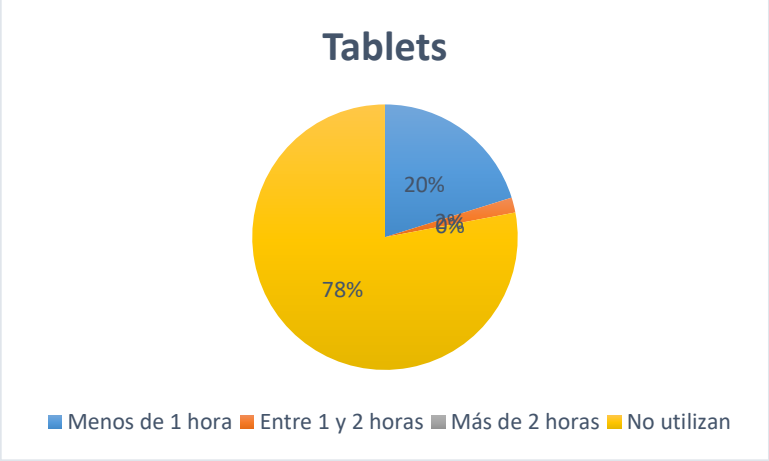


Figura 12: Porcentajes de sujetos según el tiempo de utilización de tablets

		Asiento							Total	Porcentaje
		1	2	3	4	5	6	7		
Televisión	Menos de 1 hora	2	1	8	1	4	4	4	24	22%
	Entre 1 y 2 horas	6	6	7	1	6	5	5	36	33%
	Más de 2 horas	2	8	1	7	4	6	4	32	29%
	No utilizan	5	1	0	6	2	1	2	17	16%

Tabla 17: Cantidad de observadores según el tiempo de visualización de contenido audiovisual en televisor

Figura 13: Porcentajes de sujetos según el tiempo de utilización de televisores

		Asiento							Total	Porcentaje
		1	2	3	4	5	6	7		
PC	Menos de 1 hora	6	6	4	6	7	3	3	35	32%
	Entre 1 y 2 horas	1	0	1	1	6	2	1	12	11%
	Más de 2 horas	4	3	2	4	1	4	7	25	23%
	No utilizan	4	7	9	4	2	7	4	37	34%

Tabla 18: Cantidad de observadores según el tiempo de visualización de contenido audiovisual en PC

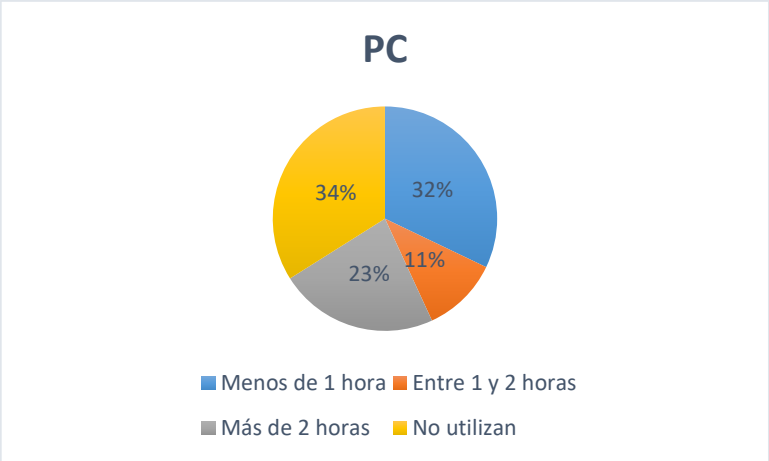


Figura 14: Porcentajes de sujetos según el tiempo de utilización de PCs

3.3 Particularidades de las sesiones de pruebas

3.3.1 Interrupciones durante las pruebas

Durante cada sesión de prueba se colocó un cartel del lado de afuera de la sala indicando que no se debía entrar, y la hora de finalización de la prueba. Esto no impidió que en algunas ocasiones alguien abriera la puerta e intentara entrar a la sala donde se llevaba a cabo la prueba. Si bien esto podría haber causado que los participantes se distrajeran, este no fue el caso.

3.3.2 Desconexión de los móviles de prueba

El segundo problema que se observó durante las pruebas, fue que a veces un móvil utilizado por un participante para votar se desconectó del access point. Esto provocó que en algún caso el participante no pudiera votar el video que acababa de ver, debido a que la aplicación se saltaba este paso al reconectar. En esos casos el voto del participante quedaba marcado como “0”. Entre las votaciones de los participantes se detectaron 8 ceros para diferentes videos y diferentes sesiones, de un total de 9156 votaciones (84 videos por 109 participantes). Dado que la cantidad de ceros fue reducida respecto al total, no fue necesario tomar ninguna acción. De hecho, quitar los votos “0”, no hace una diferencia significativa de los resultados que se exponen en el capítulo 4.

3.3.3 Problema de reproducción

Otro inconveniente que surgió durante las pruebas fue de reproducción de los videos. Para aquellos videos con tasa de bits más altas se observó que el movimiento no era continuo, sino que era entrecortado o de a saltos. Como este defecto no fue intencional y se detectó durante las pruebas, no se advirtió de su existencia a los observadores. Cada observador tomó su decisión sobre si incluirlo en su calificación del video o no. Esto en principio podría haber

producido que un observador fuera descartado porque sus calificaciones consideraron el efecto mientras los demás no lo hicieron. Dado que esto era una cuestión de preferencia individual, puesto que hay gente que decidió tomar en cuenta este factor y gente que no al momento de calificar, y gente que no notó este defecto, decidimos durante el análisis de validación de observadores incluir la correlación según HRCs que se explica en 2.4.2.2. Esto último disminuyó la cantidad de outliers.

3.3.4 Comportamiento de los observadores

En todas las sesiones, los observadores prestaron especial atención a las instrucciones provistas al inicio, tanto para las pruebas de agudeza visual y de daltonismo como para la calificación de los videos. Luego de iniciadas las sesiones de evaluación, ningún observador se movió de su asiento.

En quince de las dieciséis sesiones, los observadores permanecieron en silencio durante la calificación de los videos, interrumpiendo dicho silencio solo en caso de que se desconectara el dispositivo móvil que le fue provisto. En la sesión restante, los observadores fueron un grupo de compañeros de trabajo y hablaron entre sí hasta que se les instruyó calificar sin hacer referencia a elementos particulares de la secuencia visualizada, de modo de evitar condicionar la calificación de los otros observadores.

Al final de cada sesión, se les pidió a los observadores su opinión acerca de la sesión.

Sus opiniones se describen a continuación:

- La mayoría de los observadores encontraron aceptable la duración de la sesión.
- La mayoría de los observadores, especialmente los jóvenes, las describieron como una experiencia interesante.
- Algunos de los observadores reportan haberse aburrido al ver seis veces el mismo contenido, por más que fuera con distintos grados de degradación.
- Para PVSs que provenían de algunos SRCs en particular, los observadores parecieron encontrar difícil diferenciar grados de calidad. En el caso de las SRCs “Bund Nightscape”, “Tree Shade” y “Scarf” los observadores reportaron que sus opiniones eran siempre positivas para la mayoría de las PVSs. Por otro lado, para “Coastguard”, “Marathon” y “Runners”, fue lo opuesto.

4 Capítulo 4: Análisis de los Resultados

4.1 Validación de los observadores

De las calificaciones individuales aportadas por cada observador, se hizo para cada asiento el análisis descripto en la sección 2.4. A partir de esto, se obtuvieron para cada asiento, los valores de MOS e intervalos de confianza asociados de cada PVS y se determinó cuáles observadores serían descartados en cada asiento. Dado el volumen de datos a manejar en estos cálculos, omitimos su presentación en este documento, pudiendo consultarse el archivo "estadisticas.xls" por más información. En la Tabla 19, se muestra para cada asiento la cantidad observadores cuyas opiniones fueron validadas y la cantidad de observadores cuyas opiniones fueron descartadas.

Asiento	Observadores Validados	Observadores Descartados
1	15	0
2	14	2
3	13	3
4	9	6
5	11	5
6	15	1
7	12	3

Tabla 19: Validación de observadores

Vemos que, en la mayoría de los asientos, tenemos una cantidad de observadores validados inferior a quince, que es la requerida por la recomendación ITU-R BT.500-13 [2] para que se pueda considerar que los resultados obtenidos son estadísticamente significativos. No fue posible continuar haciendo sesiones hasta lograr cumplir con este requisito, y nuestro análisis se hizo tomando en cuenta en cada asiento las puntuaciones aportadas por los observadores validados. Por otro lado, en la recomendación ITU-T REC P.913 [4] se especifica que pueden usarse menos sujetos (entre 8 y 12) para estudios piloto. Si bien este estudio no fue inicialmente marcado como piloto, nuestros resultados son válidos para analizar tendencias.

Se destaca que de los observadores con problemas de visión y daltonismo de la Tabla 10, solamente uno resultó ser un outlier. Este observador tenía una visión inferior a 20/30 y estaba en la posición 7.

En la Tabla 19 se puede ver que para el asiento 4 hay un 40% de observadores descartados. Si bien el porcentaje es mayor que en el resto de los asientos, no se encontró una explicación para este resultado. Demográficamente esto no se explica, ya que sus edades van de 26 a 63 años, no había observadores con problemas de visión o daltonismo, los descartados son de ambos sexos y tienen diferente nivel educativo. Tampoco se notó una diferencia con respecto a la exposición a diferentes dispositivos para visualización de contenido audiovisual, o el tiempo dedicado a la visualización. En la Tabla 20 se encuentran los datos de estos 6 observadores.

Observador descartado	1	2	3	4	5	6
Sexo	F	M	M	M	M	M
Edad	39	63	61	26	33	26
Nivel educativo	Terciaria	Secundaria	Terciaria	Terciaria	Terciaria	Terciaria
Horas de teléfonos celulares	Menor a 1	Entre 1 y 2	Entre 1 y 2	Mayor a 2	Entre 1 y 2	Menor a 1
Horas de tablets	Menor a 1	0	0	0	0	0
Horas de televisión	Mayor a 2	0	0	Mayor a 2	0	Mayor a 2
Horas de PC	Menor a 1	Menor a 1	Menor a 1	Mayor a 2	0	Menor a 1

Tabla 20: Caracterización demográfica de los observadores descartados del asiento 4

A partir de los datos recabados por medio de pruebas de agudeza visual y daltonismo, y la encuesta realizada al inicio de la sesión, no existen elementos destacables en los participantes del asiento cuatro que permitan contextualizar la tasa de observadores descartados, por fuera de simples diferencias de opinión al momento de emitir votos. En el trabajo [10], se dio que, para un total de 25 participantes, hubo 10 observadores descartados (40% del total de participantes), donde tampoco parece haberse encontrado una explicación.

4.2 MOS e Intervalos de confianza

En el anexo C se muestra por cada PVS una gráfica con los siete valores de MOS correspondientes a cada asiento, y sus respectivos intervalos de confianza de 95% asociados. Esto nos da una idea inicial de la similitud entre los promedios de calificaciones aportados desde cada uno de los asientos. En particular, nos interesa la similitud entre las calificaciones promedio de cada una de las posiciones 1, 3, 4, 5, 6 y 7 con las de la posición 2. Para esto, una idea útil es ver si el intervalo de confianza asociado al MOS de un asiento dado se superpone con el intervalo de confianza asociado al MOS del asiento 2.

En la Tabla 21 se listan aquellos casos de PVSs para las cuales no hubo superposición de intervalos de confianza entre algún asiento y el asiento 2.

PVS	Posiciones sin superposición de I de C con posición 2
Campfire_Party3	6
Coastguard1	4,7
Coastguard3	7
Construcion_Field2	5,6,7
Construcion_Field3	5,6,7
Construcion_Field6	5
Construcion_Field7	5
Marathon3	5
Runners2	4,6
Rush_Hour1	5
Rush_Hour3	5
TrafficAndBuilding0	7
Tree_Shade0	5
Wood1	7
Wood3	7

Tabla 21: PVSs sin superposición de I de C respecto a la posición 2

El hecho de que no haya superposición de intervalos de confianza para los casos vistos en la Tabla 21 nos indica que la diferencia entre la calificación media de esos asientos y la calificación media del asiento 2 será estadísticamente significativa al analizarla con un test t de Student como el presentado en la sección 4.3.

4.3 Test t de Student

Se busca determinar para cada PVS, si las diferencias entre las calificaciones medias del asiento dos y cada uno de los otros son estadísticamente significativas. A estos efectos, para cada PVS se comparan las calificaciones medias entre cada asiento y el asiento dos, haciendo para cada comparación un test t de Student, como se describió en la sección 2.4.4.

Estos tests fueron realizados usando el software Microsoft Excel 2016, del cual se utilizó la función T.TEST.

La función T.TEST da como salida el p-valor y recibe los siguientes cuatro argumentos:

- Arreglo1: Primer conjunto de datos (Ej.: calificaciones del asiento i para la PVS j)
- Arreglo2: Segundo conjunto de datos (Ej.: calificaciones del asiento 2 para la PVS j)
- Colas: Cantidad de colas de la distribución a utilizar. En nuestro caso se usaron 2.
- Tipo: Entrada que indica la relación entre las varianzas de los conjuntos de datos a comparar. En nuestro caso, el test era de varianzas desiguales, por lo que se deja este valor en 3.

Una vez hecho el test, se decide:

- Rechazar la hipótesis nula si el p-valor es menor a 0,05. Esto implicaría que la diferencia entre las calificaciones medias es estadísticamente significativa.
- No rechazar la hipótesis nula si el p-valor es mayor o igual a 0,05. Esto implica que no se puede asegurar que la diferencia sea estadísticamente significativa.

Tras haber realizado los test t de Student correspondientes para todas las PVSs en todos los asientos, se obtuvieron los resultados vistos en la Tabla 22.

Asiento	**Cantidad de PVSs para las cuales se rechaza H_0**	**Cantidad de PVSs para las cuales no se rechaza H_0**
1	1	83
3	2	82
4	3	81
5	13	71
6	10	74
7	15	69

Tabla 22: Cantidad de PVSs para las cuales se rechaza H_0 y cantidades para las cuales no se rechaza

En la Tabla 23 se expresan estos mismos resultados en términos porcentuales. Se recuerda que se tienen un total de 84 PVSs.

Asiento	Porcentaje de PVSs para las cuales se rechaza H_0	Porcentaje de PVSs para las cuales no se rechaza H_0
1	1.19	98.81
3	2.38	97.62
4	3.57	96.43
5	15.48	84.52
6	11.90	88.10
7	17.86	82.14

Tabla 23: Porcentaje de PVSs para las cuales se rechaza H_0 y cantidades para las cuales no se rechaza

Las PVSs para las cuales se rechazó la hipótesis nula en cada asiento son las mostradas en la Tabla 24. Se resaltan en **negrita** aquellas PVSs para las cuales el intervalo de confianza del asiento en cuestión ni siquiera se superponía con el del asiento 2.

Asiento	PVS
1	Mobile1.mp4
3	Construction_Field3.mp4, Marathon2.mp4
4	**Coastguard1.mp4**, Runners1.mp4, **Runners2.mp4**
5	Campfire_Party4.mp4, Coastguard0.mp4, **Construction_Field2.mp4, Construction_Field3.mp4, Construction_Field6.mp4, Construction_Field7.mp4, Marathon3.mp4**, Mobile5.mp4, Runners2.mp4, Runners3.mp4, **RushHour1.mp4, RushHour3.mp4, Tree_Shade0.mp4**
6	**Campfire_Party3.mp4, Construction_Field2.mp4, Construction_Field3.mp4**, Construction_Field6.mp4, Construction_Field7.mp4, Marathon2.mp4, Marathon3.mp4, **Runners2.mp4**, Runners3.mp4, TrafficAndBuilding0.mp4
7	Campfire_Party5.mp4, **Coastguard1.mp4**, Coastguard2.mp4, **Coastguard3.mp4, Construction_Field2.mp4, Construction_Field3.mp4**, Construction_Field7.mp4, Marathon1.mp4, Mobile6.mp4, Scarf0.mp4, **TrafficAndBuilding0.mp4**, Tree_Shade0.mp4, **Wood1.mp4, Wood3.mp4**, Wood4.mp4.

Tabla 24: PVSs para los cuales se rechazó H_0

Se observa a partir de estos resultados que:

- En la primera fila los resultados fueron aproximadamente simétricos en términos porcentuales, con una cantidad máxima de PVSs para las cuales se rechazó H_0 de dos. O sea que los resultados parecen corresponderse con la intuición. A esto ayuda el hecho de que todos tienen una cantidad de observadores validados similar y cercana a 15, que es la cantidad considerada para alcanzar significación estadística según el criterio de la ITU.
- En la segunda fila, para los asientos 5,6 y 7, la cantidad de PVSs para las cuales no se rechaza la hipótesis nula fue similar en términos porcentuales, con una diferencia de aproximadamente un 3,6% (3 PVSs de 84) entre los asientos simétricos (5 y 6), y de un 5,95% (5 PVSs de 84) entre el 6 y el 7.
- El asiento 4 y el asiento 7 no se comportan de forma simétrica en términos porcentuales. Esto podría deberse por un lado a la baja cantidad de observadores validados en el asiento 4, que es menor a dos tercios del valor recomendado por la ITU, y al hecho de

que en ninguno de los asientos se llega a dicho valor, teniéndose de esta forma resultados que no son estadísticamente significativos según el criterio de la ITU.

- Para todos los asientos, la cantidad de PVSs para las cuales no se rechazó H_0 supera el 82% (69 PVSs de 84).
- Las PVSs listadas en la Tabla 21 vuelven a aparecer en la Tabla 24. O sea que se corrobora que para todas aquellas PVSs para las cuales no había habido superposición de intervalos de confianza entre un asiento dado y el dos, el test t de Student da como resultado un rechazo de la hipótesis nula. Esto se había previsto en la sección 4.2.

5 Capítulo 5: Conclusiones y trabajo futuro

En este proyecto se tuvo por objetivo analizar la viabilidad de realizar evaluaciones subjetivas de calidad de video en 4K con siete personas. Dado que no existen recomendaciones por parte de la Unión Internacional de Telecomunicaciones para 4K, se utilizaron como referencia las recomendaciones ITU-R BT.500 [2], ITU-T P.910 [3] e ITU-T P.913 [4], donde se dan lineamientos para evaluaciones subjetivas de calidad de video en SD y HD, y trabajos previos en el área de evaluaciones subjetivas en 4K, publicados en la IEEE y referenciados a lo largo de nuestro trabajo.

Para analizar la viabilidad de hacer sesiones de evaluación subjetiva de calidad de video en 4K con 7 personas, se diseñó un modelo de prueba, el cual se implementó mediante la realización de dieciséis sesiones con un total de 109 participantes. Una de ellas se realizó en el Salón 6 del edificio Los Talas del LATU, y las quince sesiones restantes tomaron lugar en la sala IEEE del Instituto de Ingeniería Eléctrica de Facultad de Ingeniería de UdelaR. En cada sesión los participantes calificaron 84 PVSs sin sonido y de 12 segundos de duración cada una. Estas PVSs se obtuvieron comprimiendo 14 SRCs a seis tasas de bits cada una con el codec HEVC. Las SRCs se obtuvieron de la base de datos de la *Universidad Jiao Tong de Shanghái* [11].

La forma de reclutamiento de los participantes fue por medio de redes de amigos, vecinos y conocidos de cada participante del proyecto. En nuestro caso, el 63,3% de los participantes fueron hombres y el 36,7% restante fueron mujeres. En cuanto al nivel educativo de los participantes el 71,6% tiene nivel terciario, el 27,5% tiene nivel de secundaria y el restante 0,9% nivel de primaria. Además, se cubrió un amplio rango de edades que va desde 15 a 75 años. Todos los observadores participaron voluntariamente de las pruebas.

Durante las sesiones, para registrar los datos de los participantes y sus votaciones se utilizó una aplicación desarrollada en un proyecto anterior, llamado VQI. A esta aplicación se le introdujeron cambios de código para que se registre la posición del observador. Esto resultó de utilidad a la hora de procesar los datos. Además, creemos que este cambio será de utilidad en proyectos futuros.

Una vez llevadas a cabo las dieciséis sesiones, se realizaron distintos tests estadísticos. El primero de ellos, consistió en analizar las votaciones y descartar a los observadores cuyas opiniones se desviaban considerablemente respecto de la media haciendo un análisis de correlación de Pearson por PVS y HRC, como se recomienda en la ITU-T P.913 [4]. Cabe destacar que en la recomendación ITU-R BT.500 [2] se sugiere tratar con un mínimo de quince observaciones válidas por cada PVS. Por otro lado, en la recomendación ITU-T REC P.913 [4] se especifica que pueden usarse menos sujetos (entre 8 y 12) para estudios piloto. Si bien este estudio no fue inicialmente marcado como piloto, nuestros resultados son válidos para analizar tendencias.

Una vez descartados estos observadores (outliers), se procedió a comparar para cada PVS las votaciones provistas desde cada una de las posiciones 1,3,4,5,6,7, con las votaciones provistas por los observadores situados en la posición 2 (posición central y usual para pruebas subjetivas con un individuo). La herramienta estadística utilizada para esta comparación fue el test t de Student para muestras con tamaño y varianza distintos y con distribución de dos colas. Dicho test sirve para determinar si las medias de dos muestras presentan diferencias estadísticamente significativas, lo cual, en nuestro caso, se traduciría a ver si existe una diferencia estadísticamente significativa entre el MOS de una PVS en un cierto asiento, y el MOS de la

misma PVS en el asiento dos. Los resultados de estos 504 tests (6 comparaciones con 84 PVSs cada una) muestran un alto grado de similitud entre las votaciones de los distintos asientos, y las del asiento dos. En términos numéricos, de las 84 comparaciones realizadas para cada uno de los asientos 1, 3, 4, 5, 6 y 7 con el asiento 2, hubo 83 (98,8%), 82 (97,6%), 81 (96,4%), 71 (84,5%), 74 (88,1%) y 69 (82,1%) comparaciones respectivamente para las cuales no se rechazaba la hipótesis nula del test t de Student. O sea que el peor caso fue el de la posición 7, para la cual en 69 de las 84 PVSs no se pudo rechazar la hipótesis de que no existe diferencia estadísticamente significativa entre las calificaciones medias. Cabe mencionar, que para aquellos casos en que se rechazó la hipótesis nula, esto se hizo con una confianza del 95%.

Concluimos entonces que la posición de los observadores en las condiciones de nuestra prueba, no influye en las calificaciones que éstos asignan a los videos presentados, y que es viable hacer sesiones de evaluación subjetiva de calidad de video en 4K con 7 personas.

Trabajo a futuro

Restaría para estudios posteriores la realización de más sesiones de evaluación subjetiva de calidad de video en 4K en las condiciones detalladas en este documento, de tal forma de alcanzar en cada asiento una cantidad de observadores validados estadísticamente significativa de acuerdo al criterio de la Unión Internacional de Telecomunicaciones. Tras obtener tal cantidad de observadores, podrían realizarse otros análisis, considerando las opiniones de todos los observadores con el mismo peso en las estadísticas del grupo, independientemente de su ubicación en la distribución espacial. Por ejemplo, se podrían realizar nuevos análisis de correlación para determinar la cantidad de observadores atípicos de la totalidad de los observadores. También, se podrían realizar análisis de las puntuaciones brindadas según distintos perfiles de observadores, determinados por nivel educativo, sexo, edad y horas de uso de distintos dispositivos para ver contenidos audiovisuales.

Podría también hacerse una modificación a la aplicación que toma las calificaciones de los observadores, de forma tal que registre el orden en que se muestran los videos en cada sesión. De esta forma, podría analizarse el efecto del cansancio en las votaciones de los participantes.

Referencias

[1] «T. K. Tan et al., “Video Quality Evaluation Methodology and Verification Testing of HEVC Compression Performance,” IEEE Trans. Circuits Syst. Video Technol., vol. 26, no. 1, pp. 76–90, Jan. 2016.».

[2] «"Methodology for the Subjective Assessment of the Quality of Television Picture", ITU-R Rec. BT.500, ITU-R, Jan 2012.».

[3] «“Subjective video quality assessment methods for multimedia applications”, Recommendation ITU-T P.910, Apr 2008.».

[4] «"Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment", Recommendation ITU-T P.913, Mar 2016.».

[5] «Y. Zhu, L. Song, R. Xie, and W. Zhang, “SJTU 4K video subjective quality dataset for content adaptive bit rate estimation without encoding,” in 2016 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2016, pp. 1–4.».

[6] «K. Berger, Y. Koudota, M. Barkowsky, and P. L. Callet, “Subjective quality assessment comparing UHD and HD resolution in HEVC transmission chains,” in 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), 2015, pp. 1–6.».

[7] «S. H. Bae, J. Kim, M. Kim, S. Cho, and J. S. Choi, “Assessments of Subjective Video Quality on HEVC-Encoded 4K-UHD Video for Beyond-HDTV Broadcasting Services,” IEEE Trans. Broadcast., vol. 59, no. 2, pp. 209–222, Jun. 2013.».

[8] «P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, “Subjective quality evaluation of the upcoming HEVC video compression standard,” Proc. SPIE, vol. 8499, 2012.».

[9] «M. Cheon and J. S. Lee, “Subjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience,” IEEE Trans. Circuits Syst. Video Technol., vol. PP, no. 99, pp. 1–1, 2017.».

[10] «Sotelo, R., Joskowicz, J., Anedda, M., Murroni, M., & Giusto, D. D. (2017). Subjective video quality assessments for 4K UHDTV. In 2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) (p. 1-6). IEEE».

[11] «L. Song, X. Tang, W. Zhang, X. Yang, P. Xia, The SJTU 4K Video Sequence Dataset, the Fifth International Workshop on Quality of Multimedia Experience (QoMEX2013), Klagenfurt, Austria, July 3rd-5th, 2013».

[12] «https://github.com/Telecommunication-Telemedia-Assessment/SITI».

[13] «https://www.ffmpeg.org/».

[14] «http://x265.readthedocs.org/en/default/».

[15] «Joskowicz, J., Sotelo, R., Juayek, M., Durán, D., & Garella, J. P. (2014, September). Automation of Subjective Video Quality Measurements. In Proceedings of the Latin America Networking Conference on LANC 2014 (p. 7). ACM.».

[16] «Joskowicz, J., Sotelo, R., Garella, J. P., Durán, D., & Juayek, M. (2015). Subjective video quality test: methodology, database and experience. In 2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (p. 1-6). IEEE».

[17] «M. H. Pinson et al., "The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study," IEEE J. Sel. Top. Signal Process., vol. 6, no. 6, pp. 640–651, Oct. 2012.».

A Anexo A: Carta de Snellen y Placas de Ishihara

A.1 Carta de Snellen

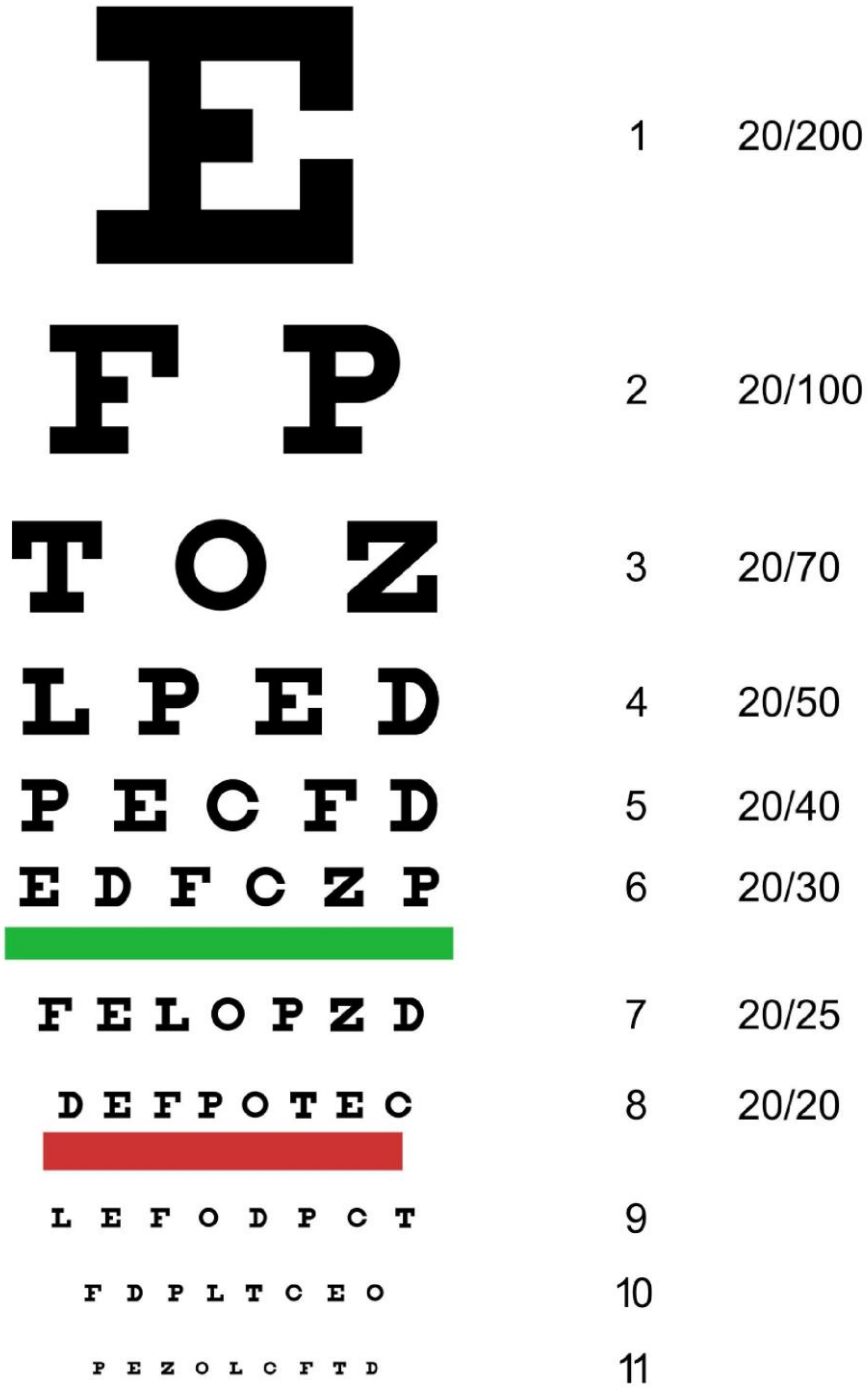


Figura 15: Carta de Snellen

A.2 Placas de Ishihara usadas

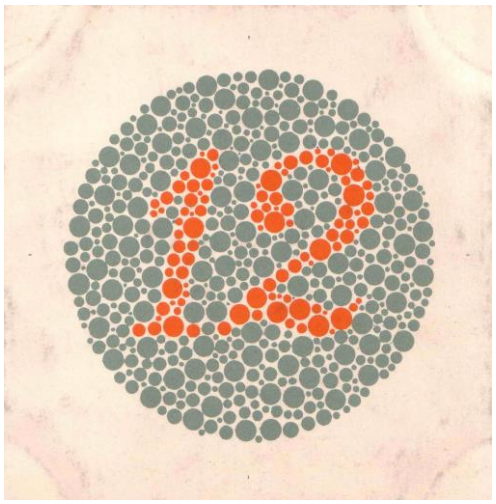


Figura 16: Placa 1. Todos ven el número 12.

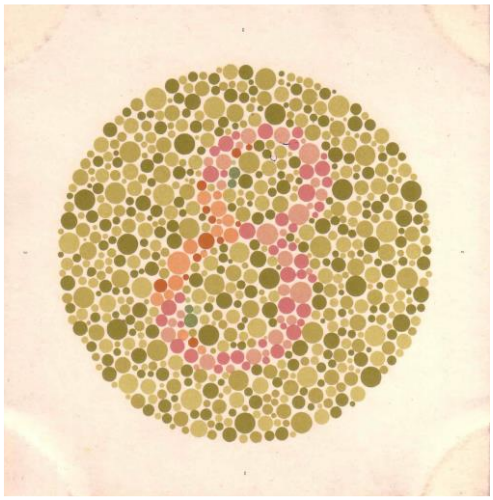


Figura 17: Placa 2. Con visión normal debe verse 8.

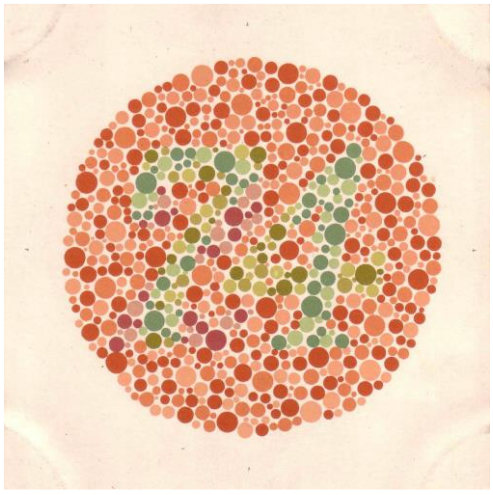


Figura 18: Placa 7. Con visión normal debe verse 74.

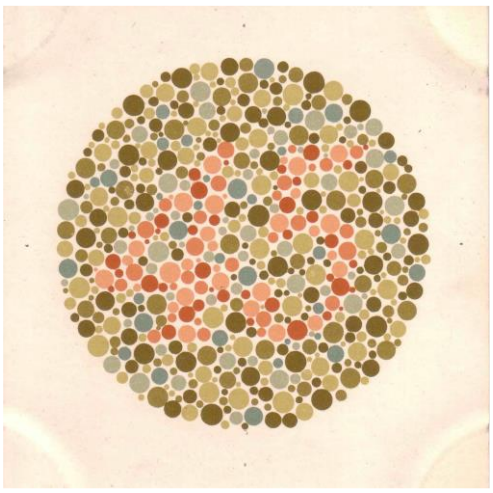


Figura 19: Placa 9. Con visión normal se ve 45.

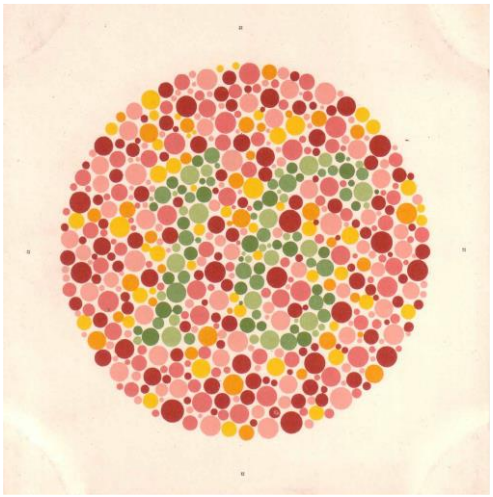


Figura 20: Placa 12. Con visión normal debe verse 16.

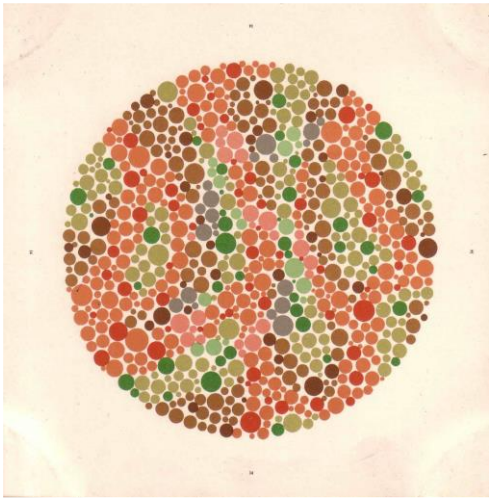


Figura 21: Placa 14. Todos ven que no hay número.

B Anexo B: Instrucciones a los observadores

Estimados participantes:

Sean bienvenidos a la prueba de evaluación de calidad de video, en el marco de nuestro proyecto de fin de carrera en la Facultad de Ingeniería de la Universidad de la República. Desde ya agradecemos su participación.

Si alguien necesita salir de la sala (por ejemplo, para ir al baño), que lo haga ahora. Una vez comenzada la prueba no se podrá abandonar su asiento hasta que la misma finalice.

Pausa para ir al baño.

Pasamos a comentarles cómo se desarrollará la prueba. Se mostrará una serie de 84 videos sin sonido de 12 segundos de duración cada uno. Después de cada video, el mismo deberá ser calificado en una escala de 5 posibles categorías:

- Excelente
- Bueno
- Aceptable
- Mediocre
- Malo

Solo se podrá calificar una vez cada video. Los mismos contenidos se verán varias veces, con diferentes tipos de degradaciones o problemas de calidad. Por favor, realice cada calificación en forma independiente de las anteriores.

Desde el navegador web de los celulares que les vamos a proveer a cada uno, completen los datos pedidos por la aplicación. El número de cédula tendrá que ingresarse sin puntos y sin el guion. El número de asiento que les pedirá la aplicación, es el que está pegado a cada una de sus sillas.

Pausa. Se les entregan los celulares a todos los usuarios.

Esta misma aplicación es la que usarán para emitir sus votos luego de cada video. En el televisor se les indicará cuándo comienza el siguiente video, y cuando el video finaliza les pedirá que lo califiquen. En ese momento aparecerán en sus celulares las 5 categorías para que califiquen. No es necesario apurarse a votar, ya que la aplicación espera a recibir todos los votos antes de pasar al siguiente video.

Pausa. También se supervisa y verifica que los usuarios completen satisfactoriamente esta etapa.

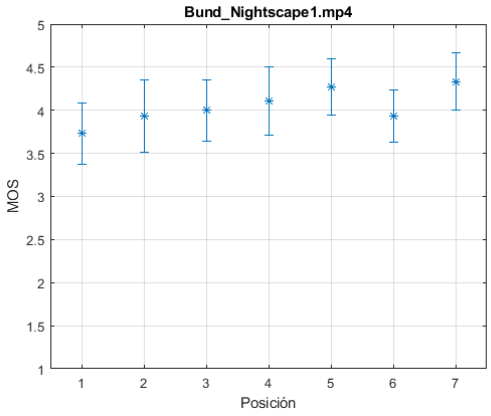
Les vamos a solicitar encarecidamente que silencien sus celulares. A aquellos que estén usando sus propios celulares para la prueba, por favor, rogamos se abstengan de usarlos con otros propósitos que no sea el de votar.

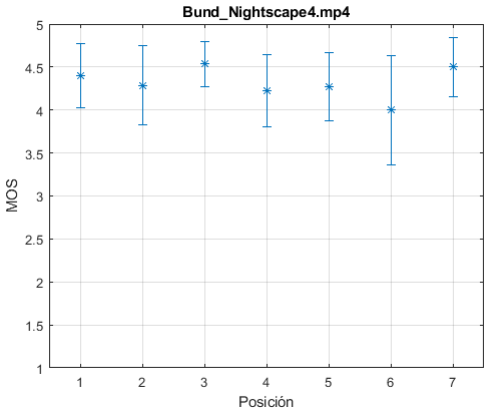
Haremos especial hincapié, en que se deberá evaluar únicamente la calidad de cada video, independientemente de su contenido, el cual puede llamar más o menos nuestra atención dependiendo de nuestros gustos.

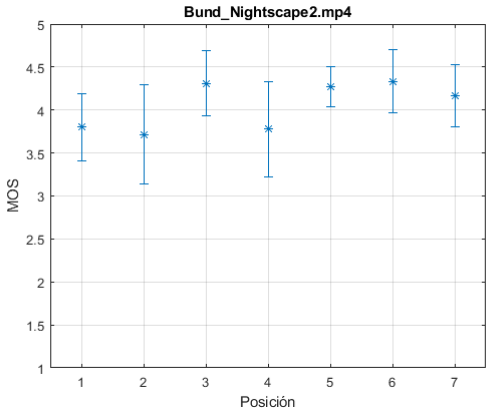
¿Alguna pregunta?

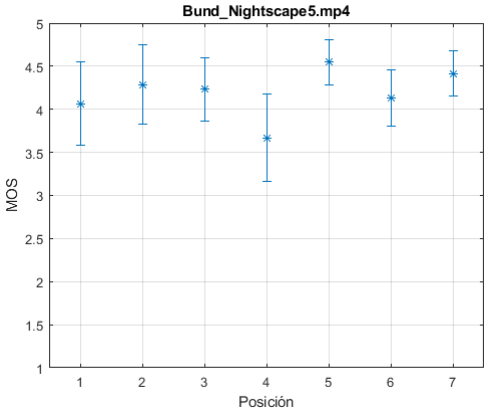
Muchas gracias.

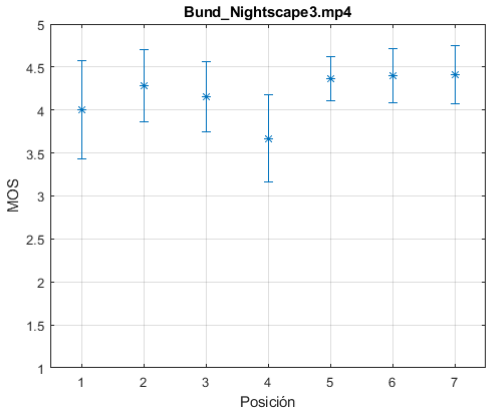
C Anexo C: Intervalos de confianza

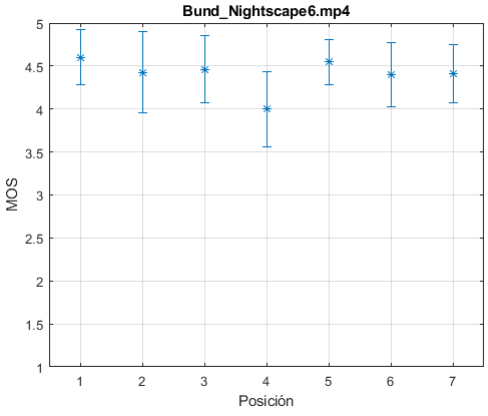


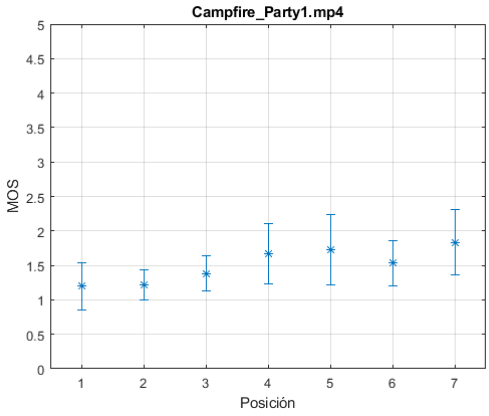


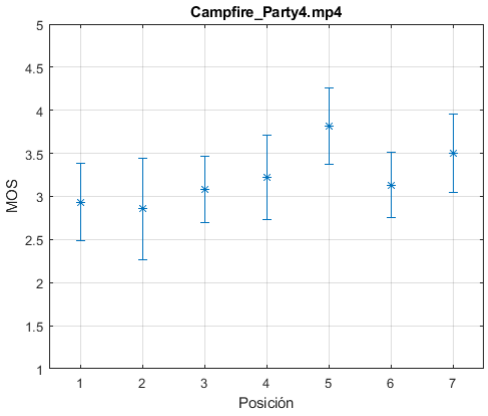


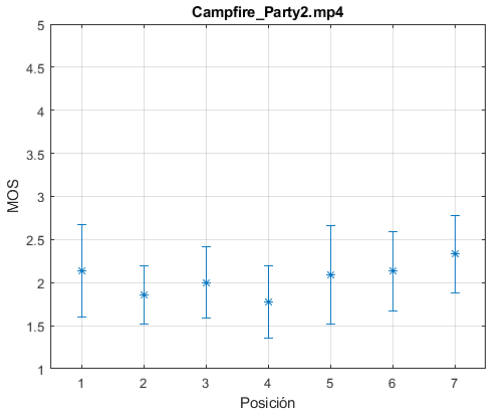


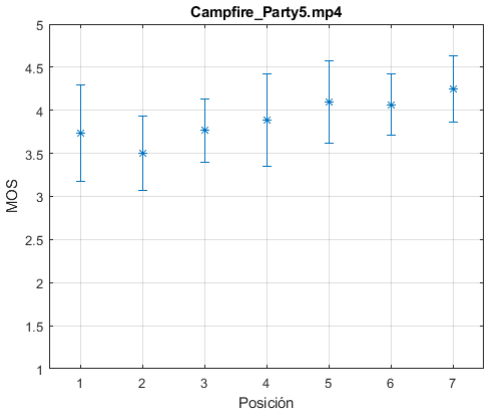


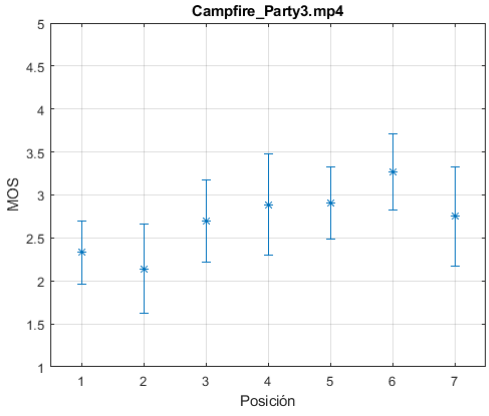


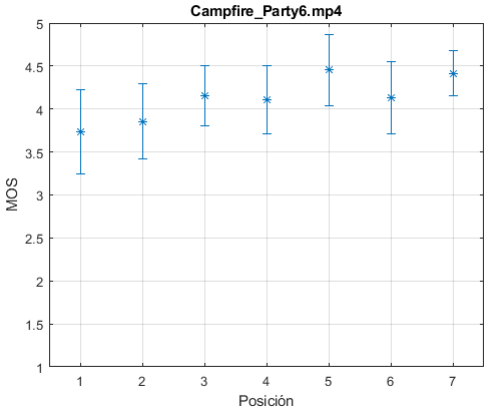

Campfire_Party1.mp4
MOS
Posición

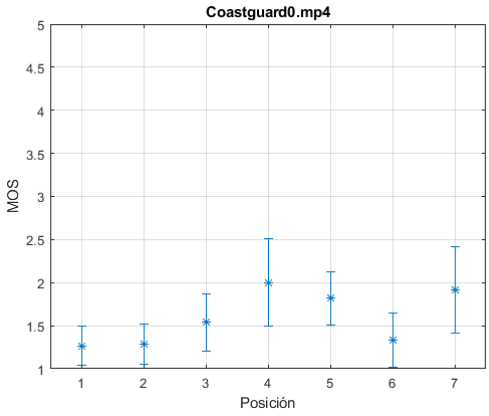

Campfire_Party4.mp4
MOS
Posición

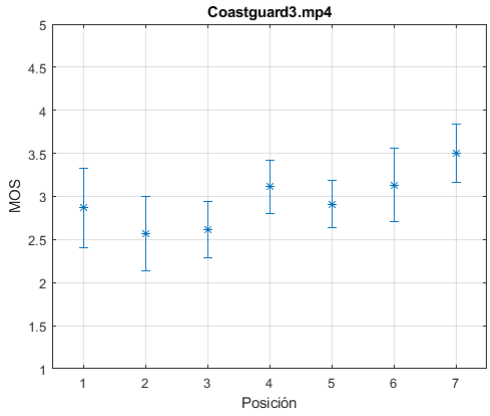

Campfire_Party2.mp4
MOS
Posición

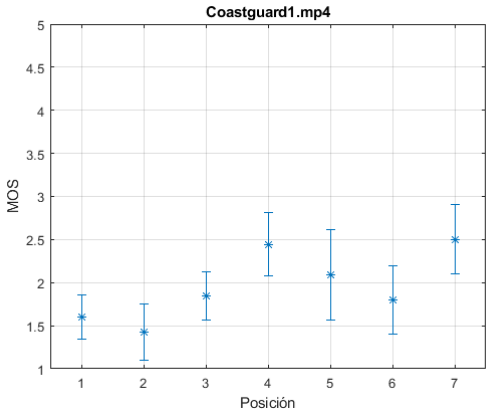

Campfire_Party5.mp4
MOS
Posición

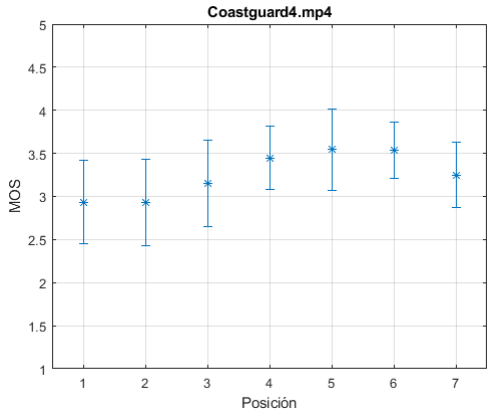

Campfire_Party3.mp4
MOS
Posición

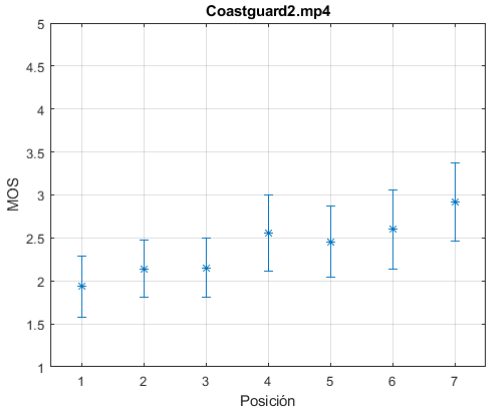

Campfire_Party6.mp4
MOS
Posición

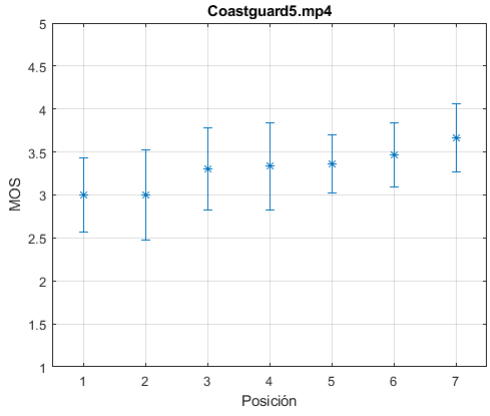

Coastguard0.mp4
MOS
Posición

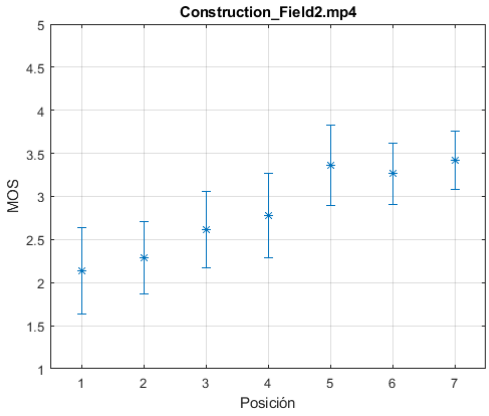

Coastguard3.mp4
MOS
Posición

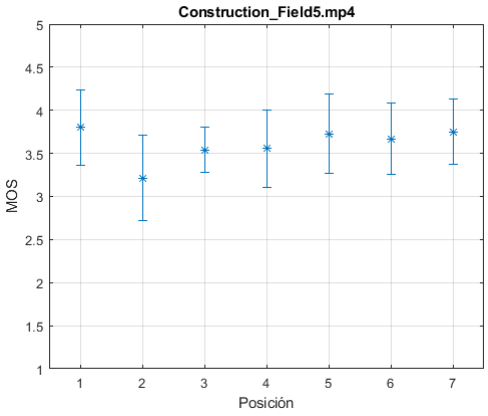

Coastguard1.mp4
MOS
Posición

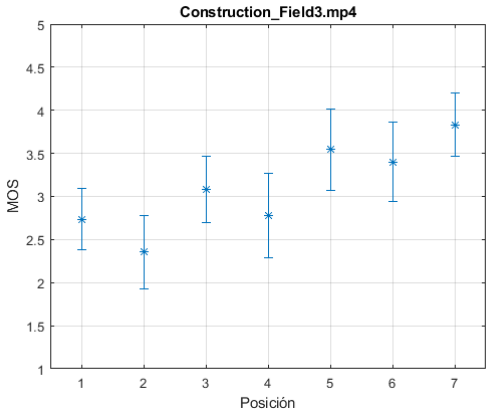

Coastguard4.mp4
MOS
Posición

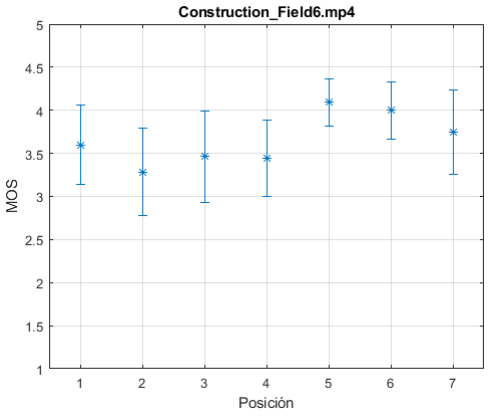

Coastguard2.mp4
MOS
Posición

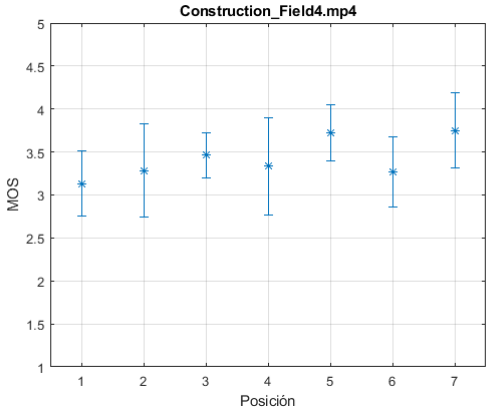

Coastguard5.mp4
MOS
Posición

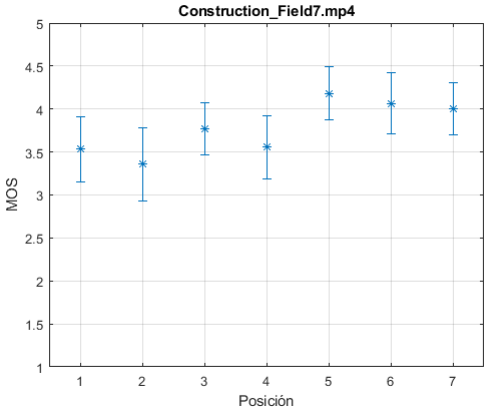

Construction_Field2.mp4
MOS
Posición

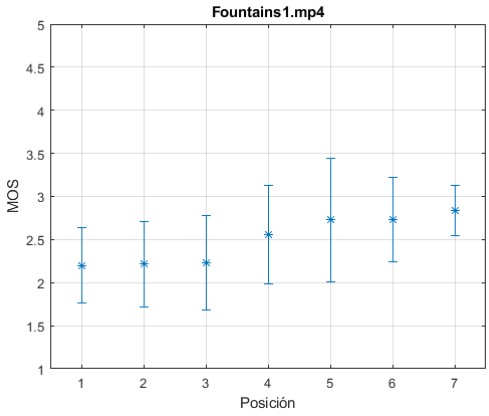

Construction_Field5.mp4
MOS
Posición

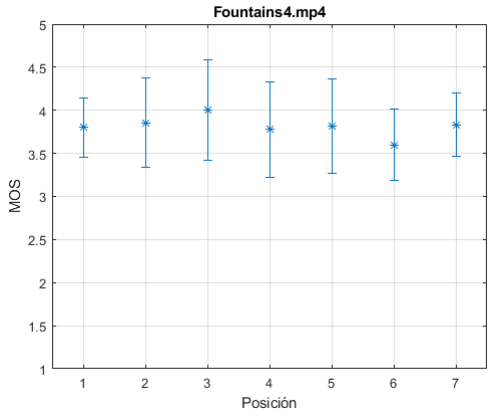

Construction_Field3.mp4
MOS
Posición

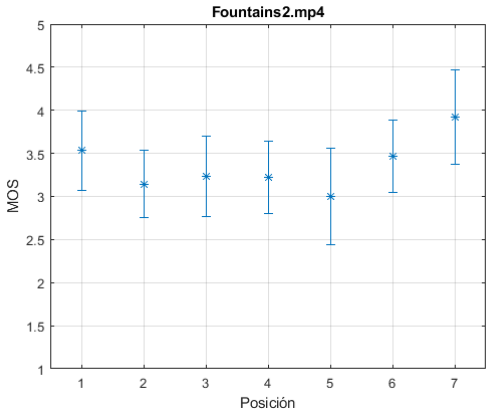

Construction_Field6.mp4
MOS
Posición

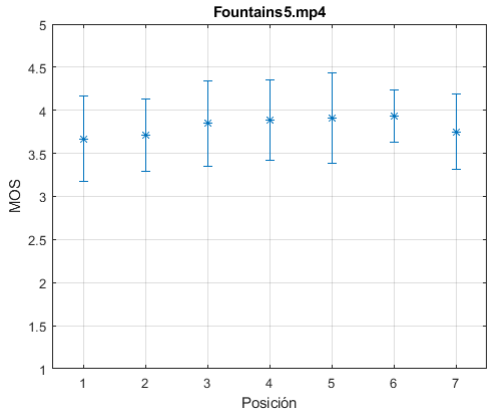

Construction_Field4.mp4
MOS
Posición

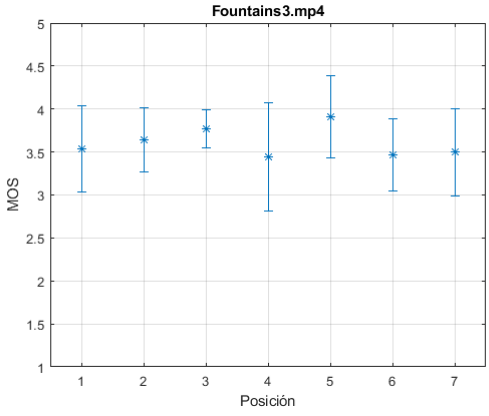

Construction_Field7.mp4
MOS
Posición

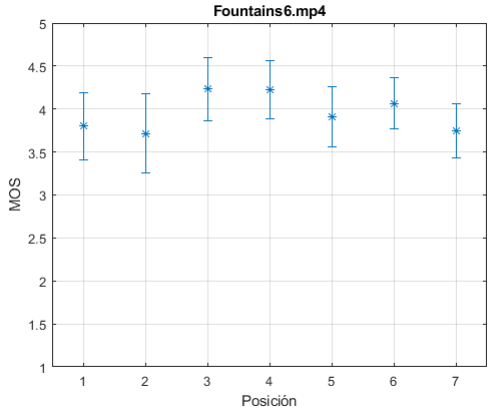

Fountains1.mp4
MOS
Posición

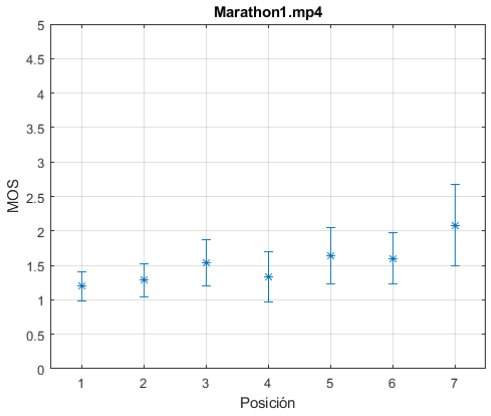

Fountains4.mp4
MOS
Posición

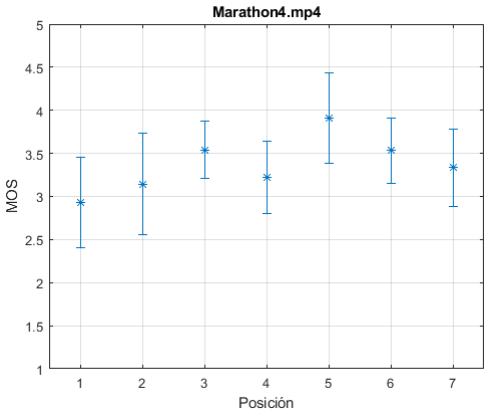

Fountains2.mp4
MOS
Posición

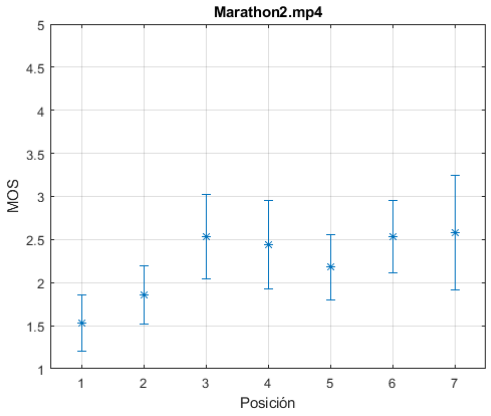

Fountains5.mp4
MOS
Posición

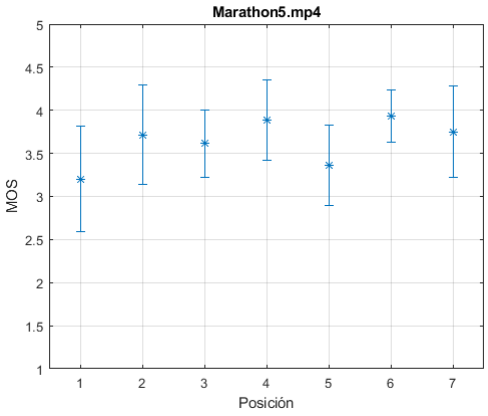

Fountains3.mp4
MOS
Posición

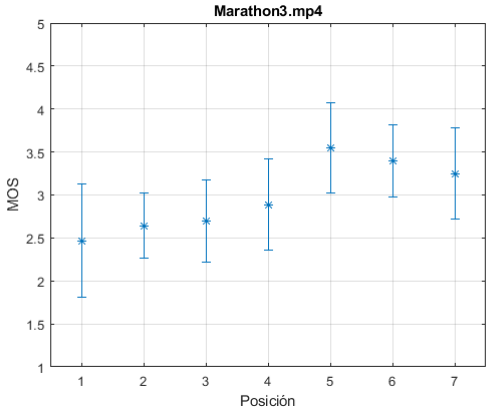

Fountains6.mp4
MOS
Posición

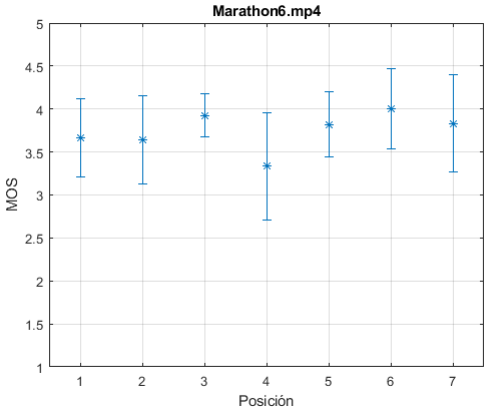

Marathon1.mp4
MOS
Posición

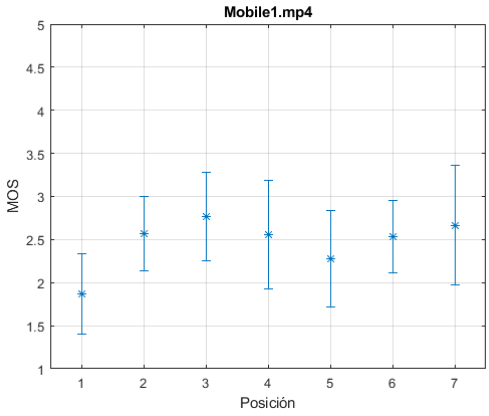

Marathon4.mp4
MOS
Posición

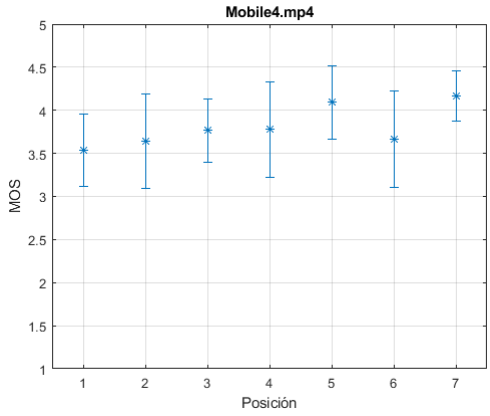

Marathon2.mp4
MOS
Posición

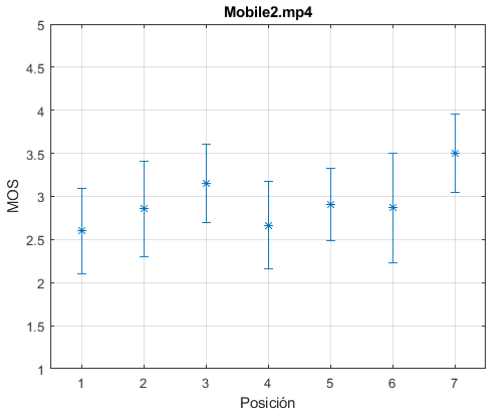

Marathon5.mp4
MOS
Posición

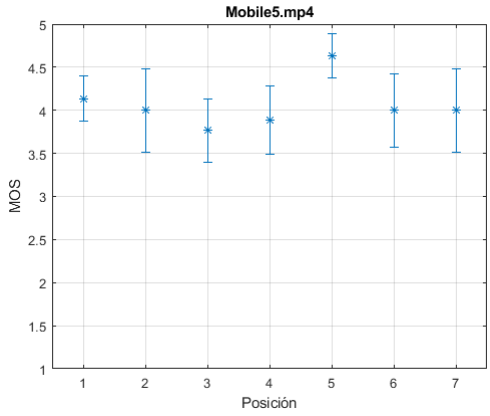

Marathon3.mp4
MOS
Posición

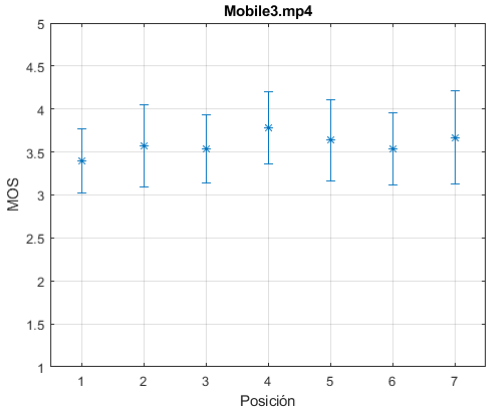

Marathon6.mp4
MOS
Posición

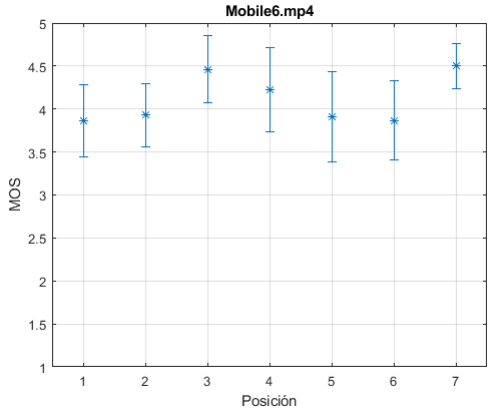

Mobile1.mp4
MOS
Posición

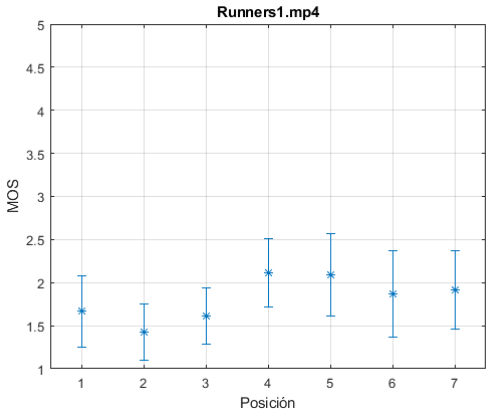

Mobile4.mp4
MOS
Posición

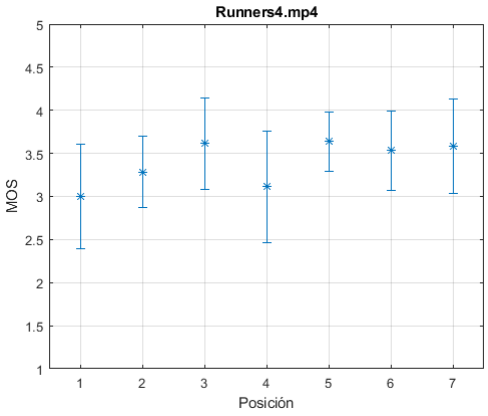

Mobile2.mp4
MOS
Posición

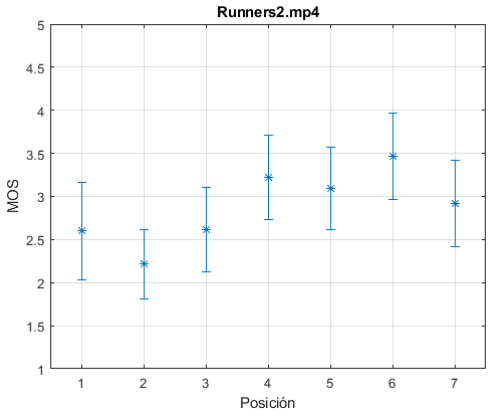

Mobile5.mp4
MOS
Posición

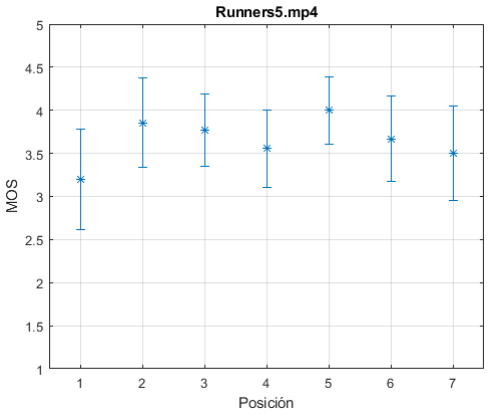

Mobile3.mp4
MOS
Posición

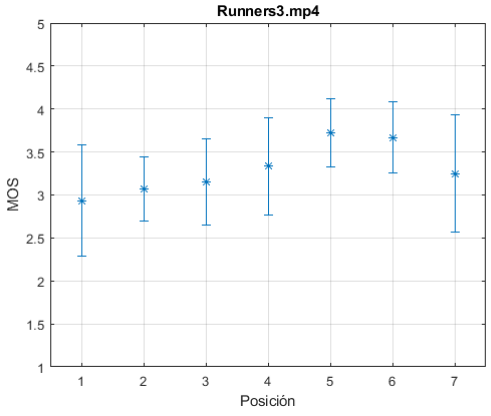

Mobile6.mp4
MOS
Posición

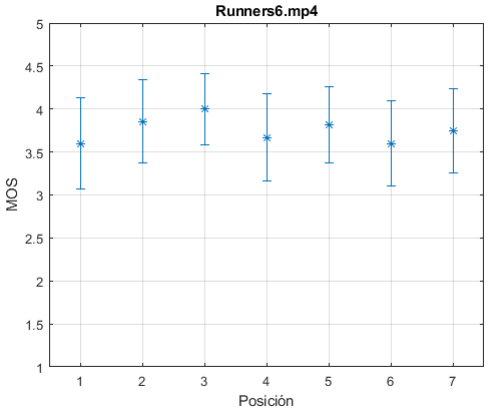

Runners1.mp4
MOS
Posición

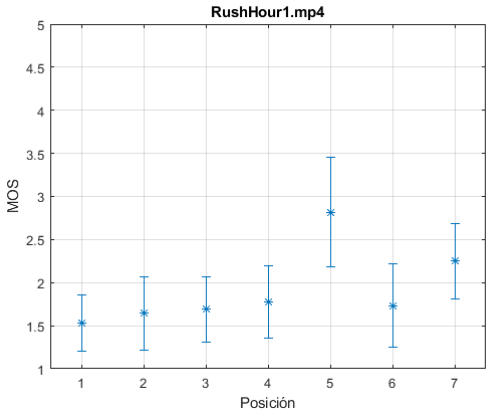

Runners4.mp4
MOS
Posición

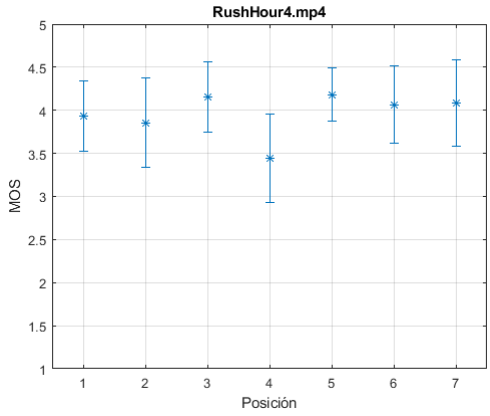

Runners2.mp4
MOS
Posición

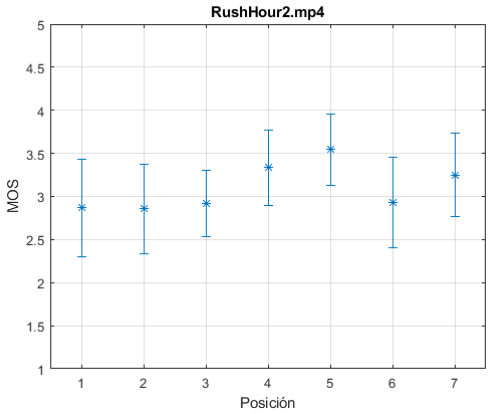

Runners5.mp4
MOS
Posición

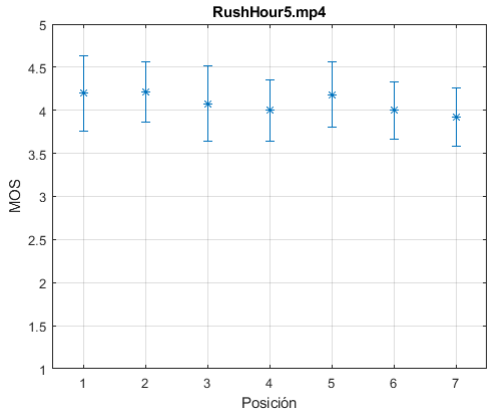

Runners3.mp4
MOS
Posición

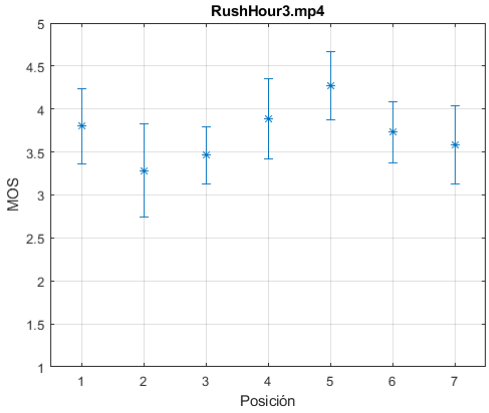

Runners6.mp4
MOS
Posición

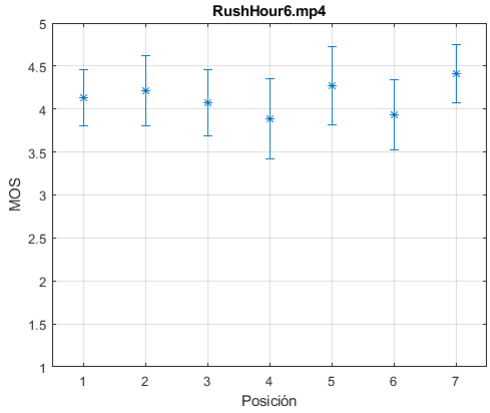

RushHour1.mp4
MOS
Posición

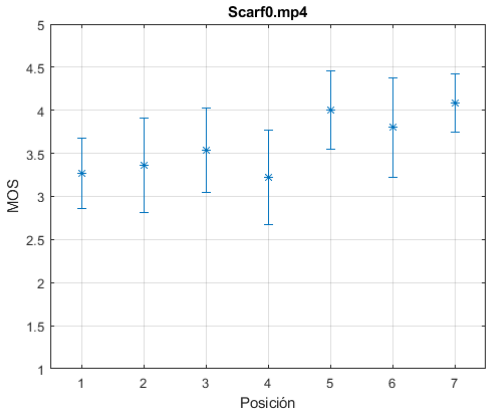

RushHour4.mp4
MOS
Posición

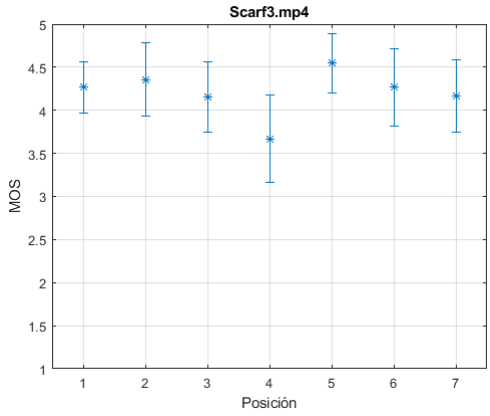

RushHour2.mp4
MOS
Posición

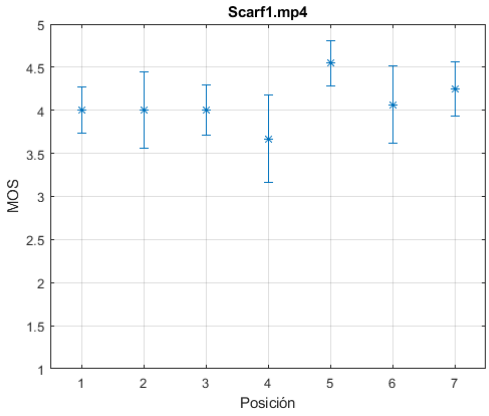

RushHour5.mp4
MOS
Posición

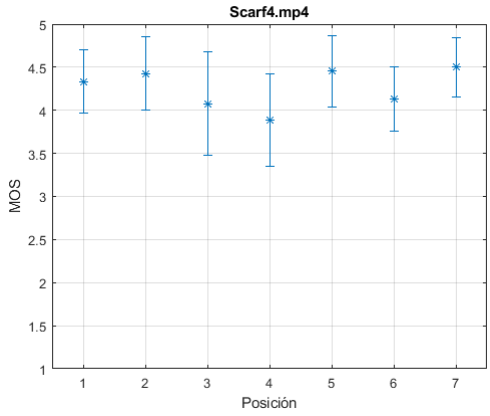

RushHour3.mp4
MOS
Posición

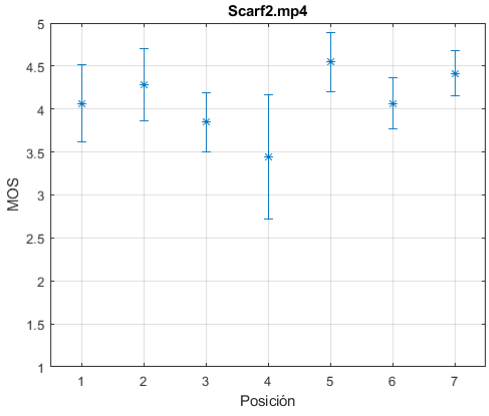

RushHour6.mp4
MOS
Posición

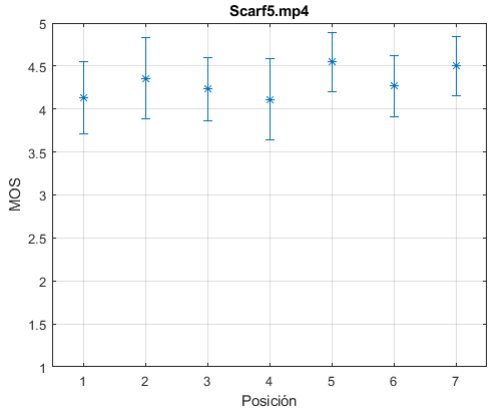

Scarf0.mp4
MOS
Posición

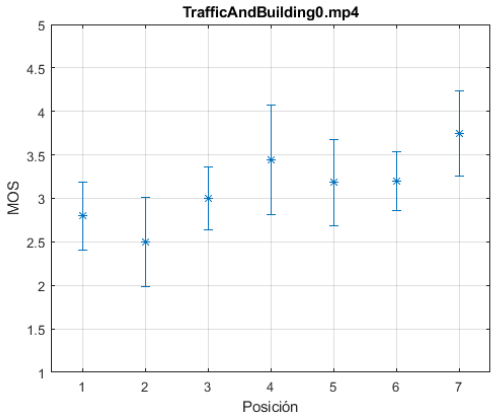

Scarf3.mp4
MOS
Posición

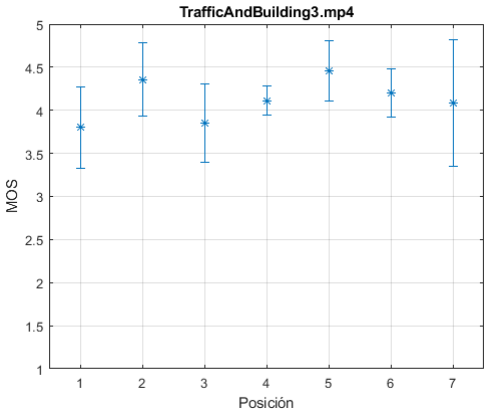

Scarf1.mp4
MOS
Posición

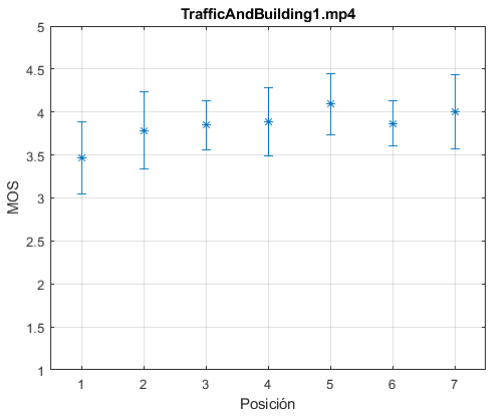

Scarf4.mp4
MOS
Posición

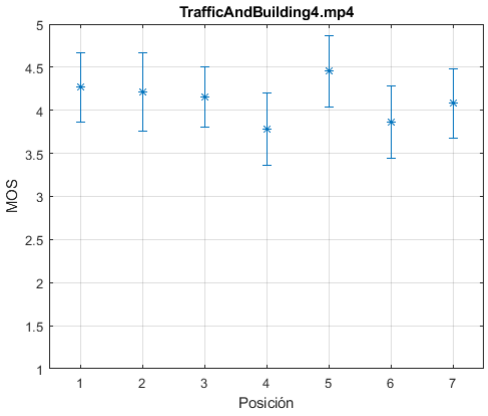

Scarf2.mp4
MOS
Posición

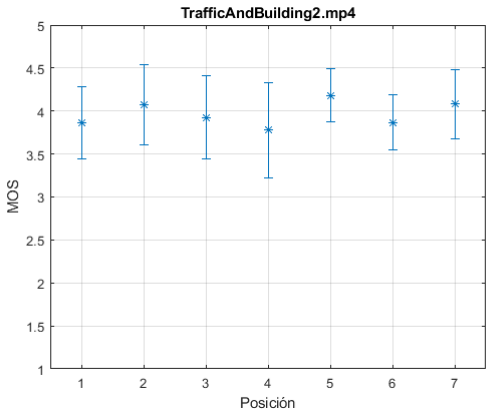

Scarf5.mp4
MOS
Posición

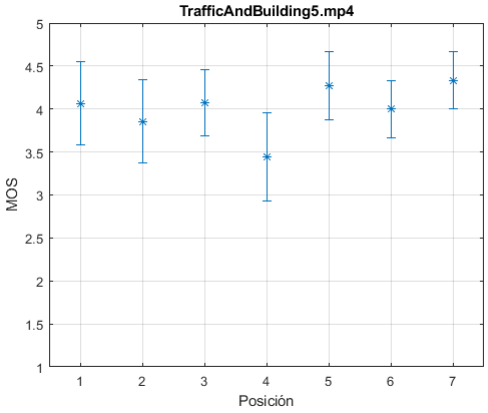

TrafficAndBuilding0.mp4
MOS
Posición

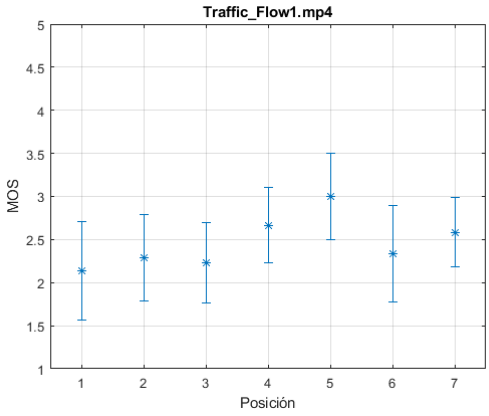

TrafficAndBuilding3.mp4
MOS
Posición

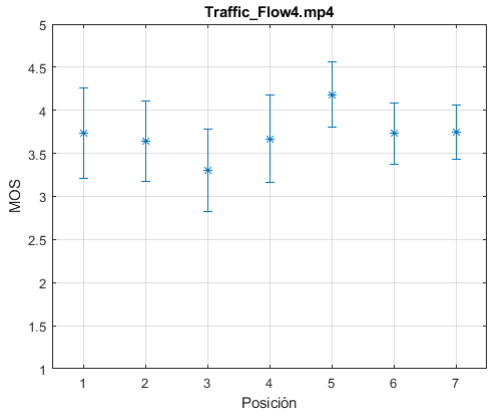

TrafficAndBuilding1.mp4
MOS
Posición

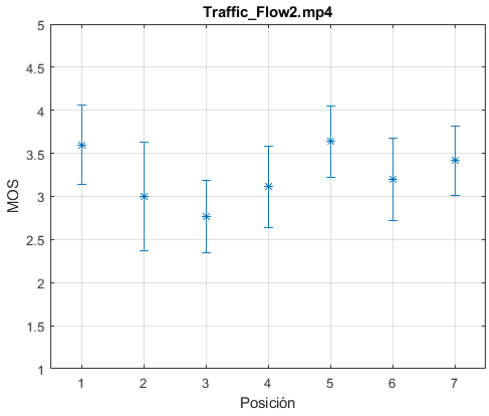

TrafficAndBuilding4.mp4
MOS
Posición

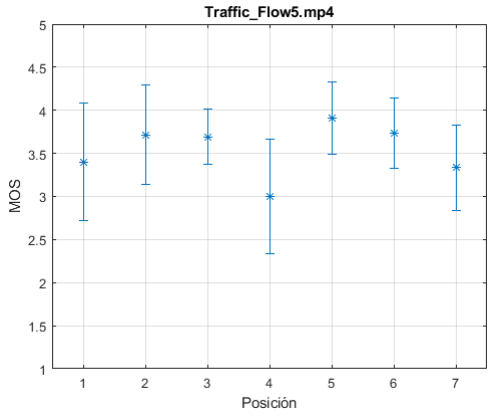

TrafficAndBuilding2.mp4
MOS
Posición

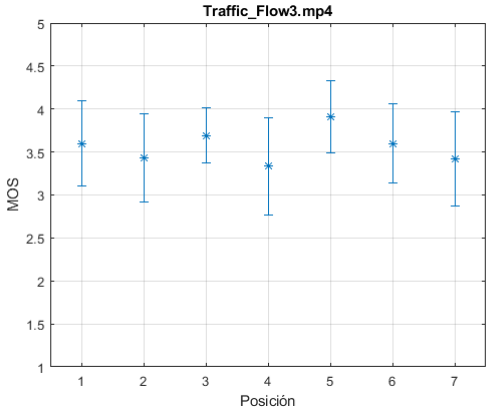

TrafficAndBuilding5.mp4
MOS
Posición

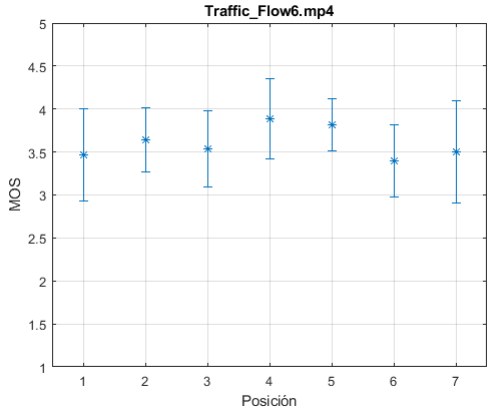

Traffic_Flow1.mp4
MOS
Posición

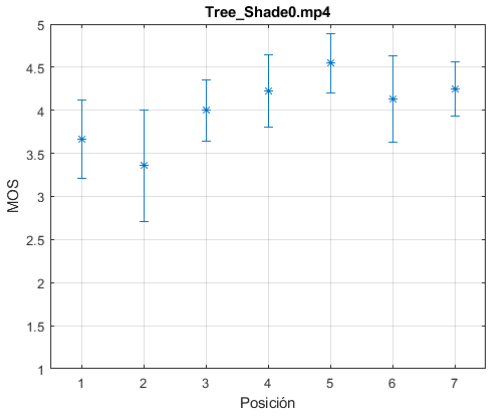

Traffic_Flow4.mp4
MOS
Posición

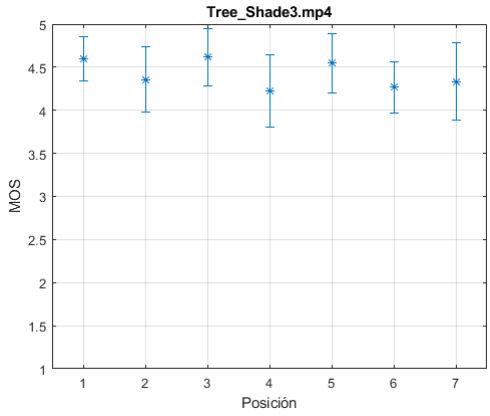

Traffic_Flow2.mp4
MOS
Posición

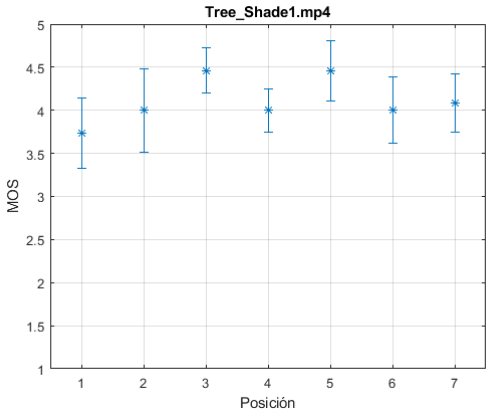

Traffic_Flow5.mp4
MOS
Posición

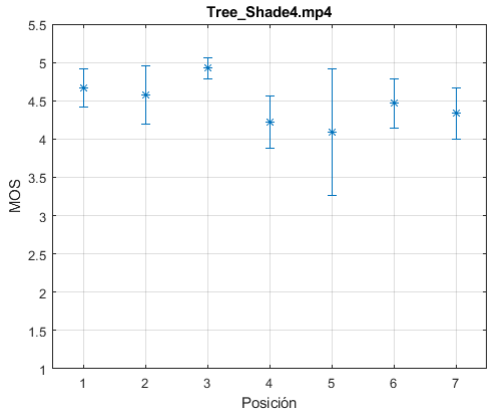

Traffic_Flow3.mp4
MOS
Posición

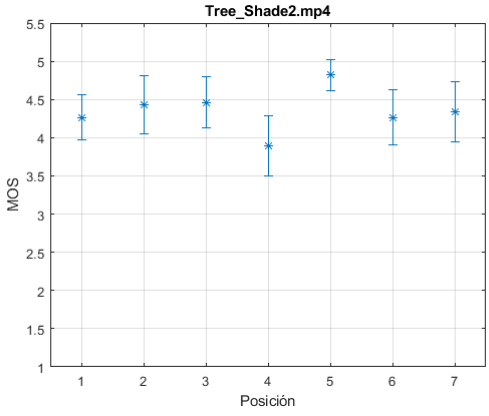

Traffic_Flow6.mp4
MOS
Posición

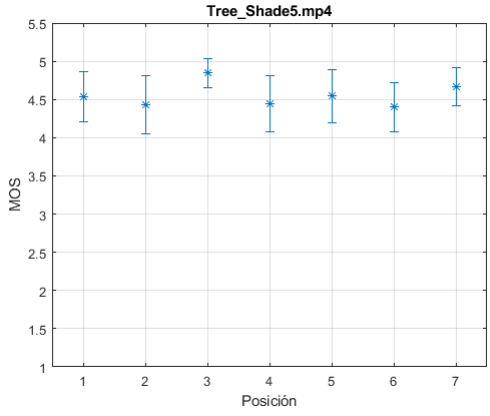

Tree_Shade0.mp4
MOS
Posición


Tree_Shade3.mp4
MOS
Posición


Tree_Shade1.mp4
MOS
Posición


Tree_Shade4.mp4
MOS
Posición


Tree_Shade2.mp4
MOS
Posición


Tree_Shade5.mp4
MOS
Posición

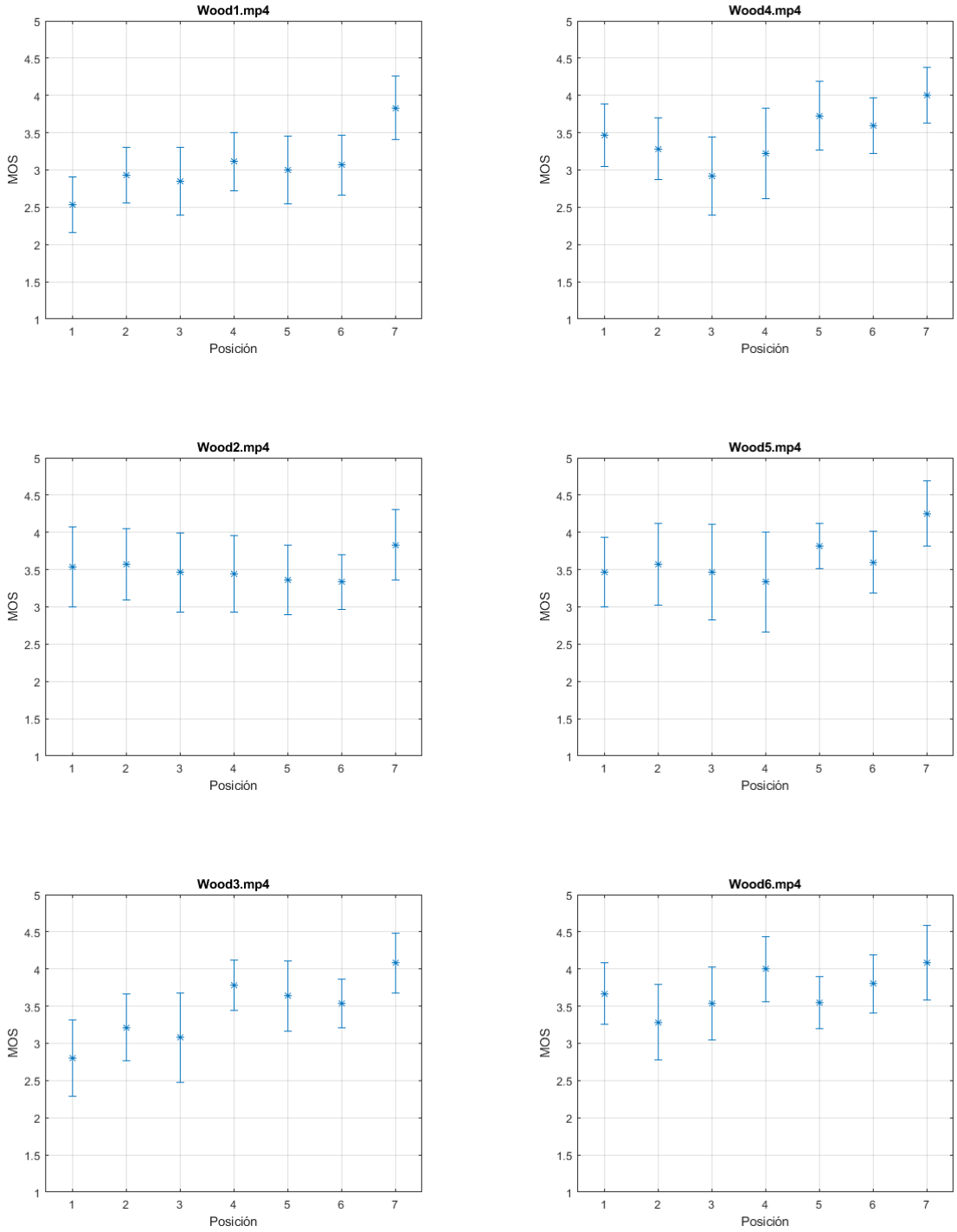


Figura 22: Intervalos de confianza para todas las PVSs

Contenido adicional

A continuación, se describen los archivos que acompañan el presente documento.

1. **plan-de-pruebas.pdf**: Archivo que contiene la versión original del plan de pruebas al 25 de octubre de 2018, previo a correcciones.
2. **datos participantes - vision y daltonismo.xlsx**: Archivo que contiene los datos de los participantes y sus resultados de los tests de visión y daltonismo.
3. **estadisticas.xlsx**: Archivo que contiene las votaciones de los participantes y los resultados de los diferentes tests realizados.
4. **pedir-asiento.zip**: Contiene el archivo nuevo y los modificados de la aplicación utilizada para recolectar los votos de los observadores. También contiene las instrucciones para la instalación.
5. **compresiones.zip**: Contiene archivos txt con los comandos utilizados para generar las PVSs a partir de las SRCs.

Plan de Pruebas Subjetivas
Proyecto QoE4k

Alejandro Goday, Federico Páramos, Laura Gomez

15 de enero de 2019

Índice

1. Introducción

En este documento se detalla el procedimiento a seguir para realizar las pruebas subjetivas de calidad en Ultra Alta Definición, así como los materiales en ellas involucrados. Esto abarca:

- Requisitos para los observadores.
- Acondicionamiento de la sala.
- Especificación del televisor.
- Distribución espacial de los observadores.
- Descripción de los videos a utilizar.
- Descripción del método de evaluación.
- Análisis de resultados

La idea general del experimento es probar con cuántos observadores se puede realizar las pruebas, sin que sus votaciones individuales presenten diferencias estadísticamente significativas. Hacer los experimentos con varias personas en lugar de con uno o dos como es usual, abarataría costos relacionados al proceso de las pruebas subjetivas, así como también reduciría el tiempo que consumen las mismas.

2. Observadores

Quienes deseen participar de la evaluación no pueden ser expertos en calidad o procesamiento de video. A todos los candidatos a observadores se les tomarán los siguientes datos:

- Cédula de identidad.
- Posición (número de asiento)
- Edad.
- Sexo (Hombre/Mujer).
- Nivel educativo.
- Cantidad de horas que mira contenidos audiovisuales en su televisor/tablet/celular/PC por día.

Al momento de realizar la prueba, los observadores recibirán instrucciones sobre la misma. Los detalles e instrucciones se dan de forma oral y se detallan en el anexo C.

Además, los candidatos deberán pasar los tests de agudeza visual y daltonismo.

2.1. Test de agudeza visual

Este test se realiza con la gráfica de Snellen. Una persona tomándolo se cubre un ojo a tres metros de distancia de la gráfica, y lee en voz alta las letras de cada fila, comenzando por la fila de más arriba. La fila más pequeña que la persona puede leer, indica con exactitud la agudeza visual en ese ojo específico. El mínimo para pasar la prueba de agudeza visual y poder participar de la evaluación es de 20/30. En caso de que el candidato use lentes habitualmente, los usará también en la prueba.

2.2. Test de Daltonismo

Los observadores son luego examinados con el test de color Ishihara, para detectar daltonismo. Éste involucra 24 (o 38) placas pseudo-isocromáticas. Cada una de ellas muestra o bien un número, o algunas líneas. Las placas se sostienen a 75cm del sujeto, y son inclinadas de forma tal que el plano del papel esté a un ángulo recto respecto a la línea de visión. Los números que se ven en las placas 1-17 son declarados, y cada respuesta debería darse sin más de tres segundos de retraso. Si el sujeto es incapaz de ver los números, se usan las placas 18-24 y las líneas onduladas entre dos cruces son rastreadas con el cepillo. Cada rastreo debería ser completado en menos de 10 segundos. No es necesario en todos los casos usar la serie completa de placas. Las placas 16 y 17 pueden ser omitidas si el test es diseñado meramente para separar la gente que tiene visión de colores defectuosa de aquellos con apreciación de color normal. En un examen a gran escala, el test puede ser simplificado al uso de solamente seis placas. Para las pruebas subjetivas se utiliza este test abreviado de seis placas. Puede ser necesario variar el orden si se sospecha que el sujeto puede intentar engañar de forma deliberada.

3. Condiciones de las pruebas

3.1. Acondicionamiento de la sala

Las pruebas se realizarán en el Laboratorio de Televisión Digital del LATU. El acondicionamiento del espacio a utilizar, seguirá las pautas vistas en la recomendación ITU-T BT.500-13 [1]. En la tabla 1, se contrastan dichas pautas con las medidas de nuestra sala.

Cuadro 1
Condiciones de Sala

Parámetros	ITU-R BT.500	Medidas de la sala
Relación entre la luminancia de pantalla inactiva y el valor de cresta de la luminancia	$\leq 0{,}02$	
Relación entre la luminancia de la pantalla, cuando sólo se muestra el nivel del negro en una sala completamente oscura, y la correspondiente al blanco más intenso	$\simeq 0{,}01$	
Brillo y contraste de la imagen	Establecido vía PLUGE	
Ángulo máximo de observación con respecto a la normal (este valor se aplica a las pantallas de tubo de rayos catódicos (TRC), para otro tipo de pantallas se están estudiando los valores adecuados):	30 grados	
Relación entre la luminancia de fondo detrás del receptor de imágenes y el valor de cresta de luminancia de la imagen	$\simeq 0{,}15$	
Cromaticidad del fondo	D_{65}	
Otra iluminación de la sala	Débil	

3.2. Especificaciones del Televisor

Para las pruebas se utiliza un televisor LED marca SONY de 65", cuyo modelo es XBR-65X755D. El mismo soporta UHD (3840 x 2160) que es condición necesaria para las pruebas.

3.3. Distribución espacial de los observadores

La disposición de los observadores en la sala es la que se indica en la figura 1.

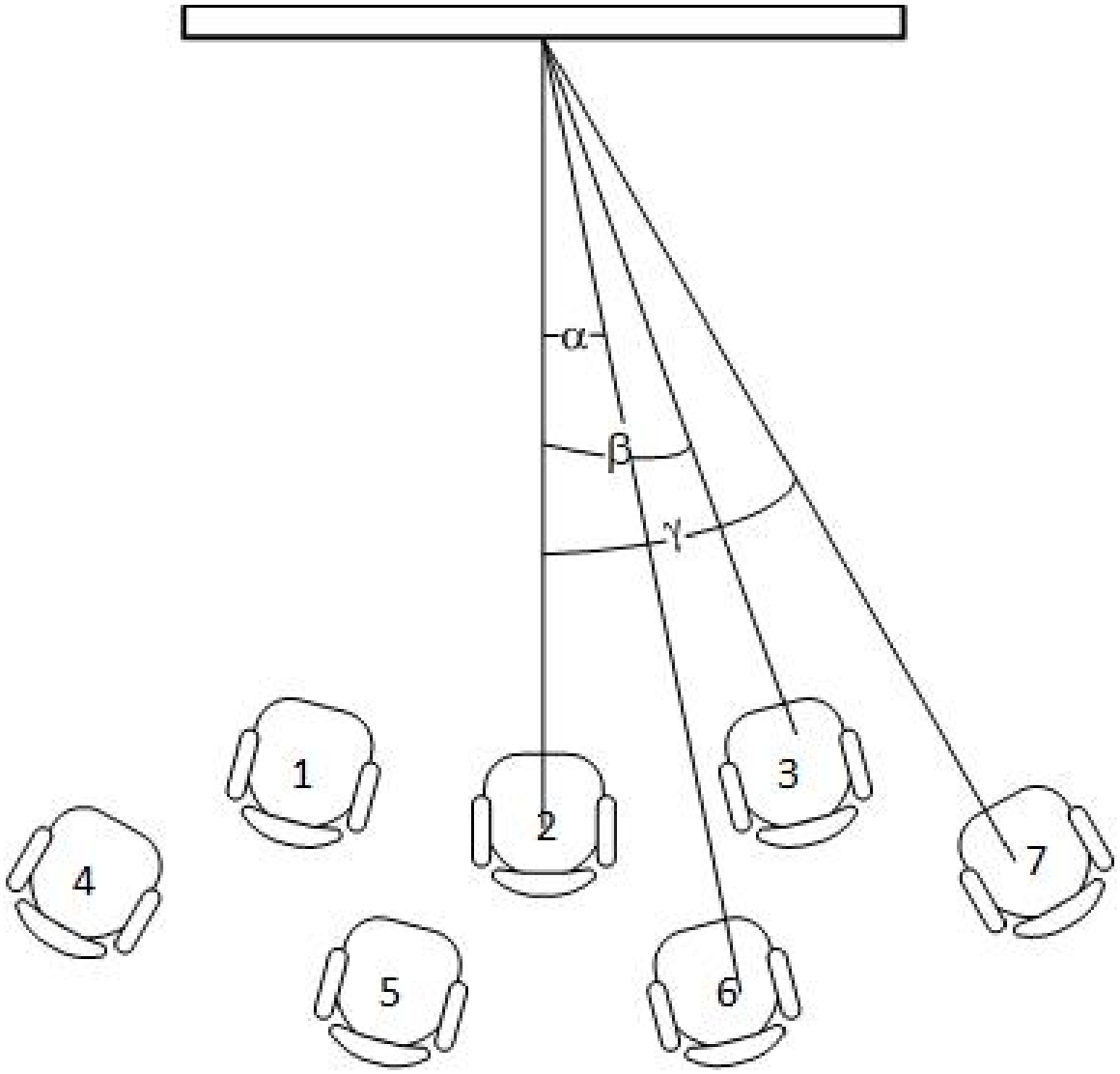


Figura 1
Distribución espacial de los observadores

La distancia d_2 será elegida en función de la altura de la pantalla como la distancia preferida de visión de acuerdo a la recomendación [1]. Por otro lado, los ángulos serán tales que el mayor de ellos (γ) respete el límite visto en la recomendación [1], o sea que habrá de ser menor a 30 grados. Los valores de las dimensiones a las cuales se hace alusión en la figura 1 se muestran en el cuadro 2.

Cuadro 2
Distribución espacial

Dimensión	Descripción	Valor
d_2	Distancia del centro de la pantalla a la posición 2	1,41 m
d_3	Distancia del centro de la pantalla a la posición 3	1,48 m
d_6	Distancia del centro de la pantalla a la posición 6	2,28 m
d_7	Distancia del centro de la pantalla a la posición 7	2,32 m
α	Ángulo entre las posiciones 2 y 6 al centro de la pantalla	14 °
β	Ángulo entre las posiciones 2 y 3 al centro de la pantalla	40 °
γ	Ángulo entre las posiciones 2 y 7 al centro de la pantalla	28 °

4. Secuencias de video originales (SRCs)

4.1. Requerimientos

A efectos de realizar las pruebas subjetivas, es necesario seleccionar secuencias de video "originales", llamadas SRC (Source Reference). Dichas secuencias deben cumplir los siguientes requerimientos:

- No deben tener degradaciones visibles a criterio del grupo QoE4k
- El origen de la secuencia debe ser conocido y registrado
- Deben tener una duración entre 10 y 12 segundos
- En lo posible, deben cubrir diferentes rangos de actividad espacial y temporal

Se evaluaron veinticuatro posibles secuencias originales, de las cuales se seleccionaron diecisiete. Se pueden ver detalles de las veinticuatro secuencias en los anexos A y B. De las diecisiete secuencias seleccionadas, solo catorce se utilizan para las pruebas subjetivas, para que la prueba total no exceda los 45 minutos. Las siete secuencias restantes fueron descartadas por excesivo ruido. Todas las secuencias evaluadas son UHD (3840 × 2160).

4.2. Complejidad espacial-temporal

Para cuantificar la complejidad espacial y temporal se han utilizado las siguientes métricas:

- Índice de información espacial (SI)
- Índice de información temporal (TI)

Se procuró seleccionar las secuencias de forma tal que se abarcasen distintos grados de complejidad espacial y temporal.

La figura 2 es una gráfica de SI contra TI donde cada punto corresponde a una secuencia del conjunto de candidatas. Los valores de SI y TI se obtuvieron con el programa SITI-master [2], y se detallan en el Cuadro 3

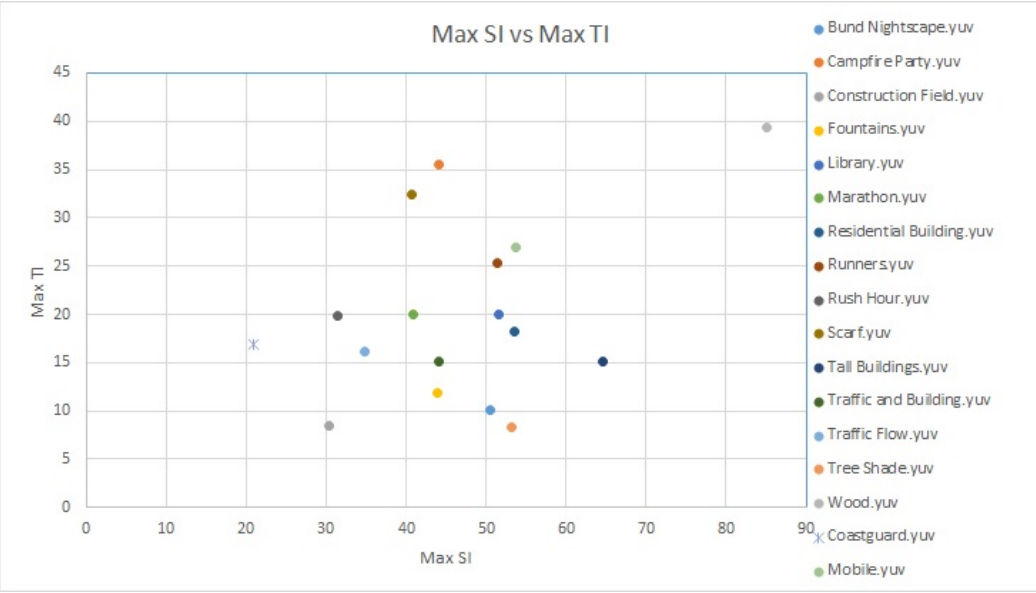


FIGURA 2
Gráfico de TI vs SI. Se muestran las complejidades espaciales y temporales para todas las secuencias candidatas.

Cuadro 3
Valores de SI y TI

Nombre del video	Origen	Max SI	Max TI
Bund Nightscape.yuv	SJTU	50,44	10,06
Campfire Party.yuv	SJTU	44,07	35,55
Construction Field.yuv	SJTU	30,39	8,45
Fountains.yuv	SJTU	43,93	11,82
Library.yuv	SJTU	51,64	20,03
Marathon.yuv	SJTU	40,85	19,05
Residential Building.yuv	SJTU	53,62	18,30
Runners.yuv	SJTU	51,40	25,35
Rush Hour.yuv	SJTU	31,43	19,91
Scarf.yuv	SJTU	40,72	32,44
Tall Buildings.yuv	SJTU	64,59	15,19
Traffic and Building.yuv	SJTU	44,04	15,13
Traffic Flow.yuv	SJTU	34,89	16,22
Tree Shade.yuv	SJTU	53,15	8,38
Wood.yuv	SJTU	85,16	39,38
Coastguard.yuv	Elemental	20,92	16,92
Mobile.yuv	Elemental	53,70	26,91

4.3. Vector de Movimiento (MV) y Suma de Diferencias Absolutas (SAD)

A efectos de cuantificar el movimiento presente en un video, es también posible calcular la magnitud de los vectores de movimiento (**MV** por Motion Vector) y la suma de diferencias absolutas (**SAD** por Sum of Absolute Differences) [4].

Para calcular esto se utilizan las siguientes herramientas: FFmpeg [5], Avisynth v.2.60 [6], Plugin MVTools (v.2.5.11.22) para Avisynth [7], VirtualDub v.1.10.4 [8], MVandSADReader.class (Programa JAVA provisto por equipo VQI).

Cálculo

1. Se usa FFmpeg [5] para pasar cada uno de los videos que tenemos con extensión YUV (<entrada>.yuv) a AVI (<entrada>.avi). A estos efectos se usa el comando:

   ```
   ffmpeg -pix_fmt yuv420p -s 3840x2160 -i <entrada>.yuv -vcodec rawvideo
   -pix_fmt yuv420p -r 30 <entrada>.avi
   ```

2. Luego, para cada video se crea un script de Avisynth (archivo de texto con extensión .avs) con la forma siguiente:

   ```
   clip = AviSource(<entrada>.avi)
   clip2 = ConvertToYV12(clip)
   clip3 = MSuper(clip2)
   vectors = MAnalyse(clip3,isb = false, outfile = "<salida>", blksize = 8)
   MShow(clip3, vectors)
   ```

3. Una vez que tenemos el script hecho, abrimos el programa VirtualDub, vamos a `File` → `Open Video File`, y ahí abrimos el script correspondiente al video cuyos MV y SAD queremos hallar. A continuación presionamos el botón de reproducir. Al finalizar la reproducción se genera un archivo binario (sin extensión), cuyo nombre es "<salida>" (como aparece en la cuarta línea del script correspondiente).

4. Para poder pasar la información de dicho archivo a formato .txt se utiliza el programa *MVandSADReader.class*.

Resultados

La figura 3 es una gráfica de MV promedio contra SAD promedio, donde cada punto corresponde a una secuencia del conjunto de candidatas. Los valores de MV promedio y SAD promedio, se obtuvieron con el a partir de los .txt generados, y se detallan en el cuadro 4.

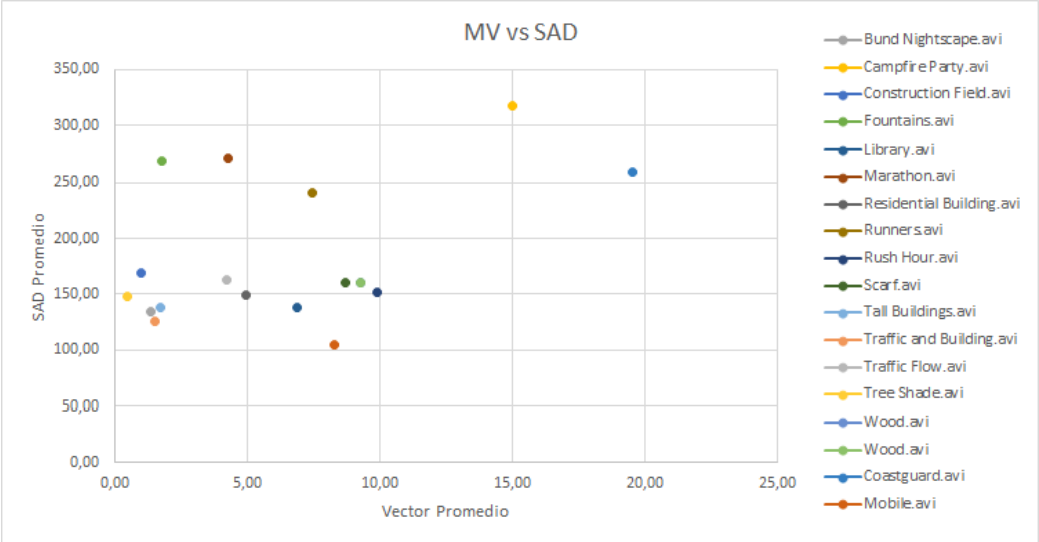


FIGURA 3
Gráfico de MV vs SAD. Se muestran el promedio de los vectores de movimiento y la suma de diferencias absolutas promedio para todas las secuencias candidatas.

CUADRO 4
Valores de MV y SAD

Nombre del video	Origen	Vector Promedio	SAD Promedio
Bund Nightscape.avi	SJTU	1,35	134,12
Campfire Party.avi	SJTU	15,00	317,10
Construction Field.avi	SJTU	0,99	168,80
Fountains.avi	SJTU	1,75	268,81
Library.avi	SJTU	6,88	138,10
Marathon.avi	SJTU	4,25	272,03
Residential Building.avi	SJTU	4,95	149,53
Runners.avi	SJTU	7,45	240,26
Rush Hour.avi	SJTU	9,89	151,44
Scarf.avi	SJTU	8,68	159,82
Tall Buildings.avi	SJTU	1,73	138,58
Traffic and Building.avi	SJTU	1,50	126,18
Traffic Flow.avi	SJTU	4,22	162,25
Tree Shade.avi	SJTU	0,48	147,70
Wood.avi	SJTU	9,26	160,04
Coastguard.avi	Elemental	19,51	259,80
Mobile.avi	Elemental	8,28	105,04

5. Secuencias de video procesadas (PVS)

Para la realización de pruebas subjetivas se generan secuencias de video degradadas, partiendo de los videos originales (SRC) y procesándolos según los distintos circuitos de referencia (HRC) establecidos. Las secuencias de video obtenidas se denominan PVS (Processed Video Sequence).

A estos efectos, se utiliza el programa *FFmpeg* [5], al cual se le dio como entrada un video con formato raw (YUV), y el cual dio como salida un archivo mp4 codificado en HEVC según el circuito de referencia utilizado. Dicho programa se ejecuta por línea de comandos y los comandos utilizados tienen la siguiente forma:

```
ffmpeg.exe -f rawvideo -video_size 3840x2160 -pixel_format yuv420p -i <entrada>.yuv
-c:v hevc -b:v <bitRate>k -maxrate <bitRate>k -minrate <bitRate>k -r 30 -x265-params
-an -y <salida>.mp4
```

Donde para cada entrada, se utilizaron los valores de Bit Rate de los cuadros 5, 6 y 7, así obteniendo seis versiones con distintos grados de degradación.

Cuadro 5
Valores de Bitrates

Salida	Nivel de Tasa	Bit Rate (kbps)
Bund_Nightscape1.mp4	Muy bajo	2341.39
Bund_Nightscape2.mp4	Bajo	5069.66
Bund_Nightscape3.mp4	Medio	7964.73
Bund_Nightscape4.mp4	Medio alto	11994.09
Bund_Nightscape5.mp4	Alto	18011.08
Bund_Nightscape6.mp4	Muy alto	24072.74
Campfire_Party1.mp4	Muy bajo	1908
Campfire_Party2.mp4	Bajo	3566
Campfire_Party3.mp4	Medio	5916
Campfire_Party4.mp4	Medio alto	7817
Campfire_Party5.mp4	Alto	12300.38
Campfire_Party6.mp4	Muy alto	16231.04
Coastguard0.mp4	Muy bajo	1491.21
Coastguard1.mp4	Bajo	1996.5
Coastguard2.mp4	Medio	3498.1
Coastguard3.mp4	Medio alto	5956.84
Coastguard4.mp4	Alto	9886.47
Coastguard5.mp4	Muy alto	11882.27
Construction_Field2.mp4	Muy bajo	927.59
Construction_Field3.mp4	Bajo	1432.58
Construction_Field4.mp4	Medio	2420.89
Construction_Field5.mp4	Medio alto	3917.54
Construction_Field6.mp4	Alto	9869.63
Construction_Field7.mp4	Muy alto	14877.93
Fountains1.mp4	Muy bajo	2404.09
Fountains2.mp4	Bajo	5240.66
Fountains3.mp4	Medio	8167.25
Fountains4.mp4	Medio alto	12190.31
Fountains5.mp4	Alto	18221.92
Fountains6.mp4	Muy alto	24236.71

Cuadro 6
Valores de Bitrates cont.

Salida	Nivel de Tasa	Bit Rate (kbps)
Marathon1.mp4	Muy bajo	2054.05
Marathon2.mp4	Bajo	4056.19
Marathon3.mp4	Medio	7514.98
Marathon4.mp4	Medio alto	11057.18
Marathon5.mp4	Alto	15046.79
Marathon6.mp4	Muy alto	18043.88
Mobile1.mp4	Muy bajo	821.53
Mobile2.mp4	Bajo	1111.46
Mobile3.mp4	Medio	1788.37
Mobile4.mp4	Medio alto	2323.05
Mobile5.mp4	Alto	5377.13
Mobile6.mp4	Muy alto	8447.09
Runners1.mp4	Muy bajo	2124.51
Runners2.mp4	Bajo	5354.24
Runners3.mp4	Medio	6753.5
Runners4.mp4	Medio alto	9930.42
Runners5.mp4	Alto	12320.76
Runners6.mp4	Muy alto	18263.62
Rush_Hour1.mp4	Muy bajo	1573.02
Rush_Hour2.mp4	Bajo	2737.13
Rush_Hour3.mp4	Medio	4798.39
Rush_Hour4.mp4	Medio alto	5988.94
Rush_Hour5.mp4	Alto	8126.02
Rush_Hour6.mp4	Muy alto	12287.15
Scarf0.mp4	Muy bajo	377.06
Scarf1.mp4	Bajo	1537.41
Scarf2.mp4	Medio	2886.03
Scarf3.mp4	Medio alto	5111.91
Scarf4.mp4	Alto	7025.3
Scarf5.mp4	Muy alto	11723.19
TrafficAndBuilding0	Muy bajo	1087.67
TrafficAndBuilding1	Bajo	2256.33
TrafficAndBuilding2	Medio	4966.11
TrafficAndBuilding3	Medio alto	7845.67
TrafficAndBuilding4	Alto	11828.24
TrafficAndBuilding5	Muy alto	17838.98
Traffic_Flow1.mp4	Muy bajo	899.09
Traffic_Flow2.mp4	Bajo	1959.23
Traffic_Flow3.mp4	Medio	4009.65
Traffic_Flow4.mp4	Medio alto	5000.24
Traffic_Flow5.mp4	Alto	7975.35
Traffic_Flow6.mp4	Muy alto	11850.27
Tree_Shade0.mp4	Muy bajo	1019.83
Tree_Shade1.mp4	Bajo	2218.25
Tree_Shade2.mp4	Medio	4871.96
Tree_Shade3.mp4	Medio alto	7627.09
Tree_Shade4.mp4	Alto	11475.88
Tree_Shade5.mp4	Muy bajo	17325.7

Cuadro 7
Valores de Bitrates cont.

Salida	Nivel de Tasa	Bit Rate (kbps)
Wood1.mp4	Muy bajo	1779.16
Wood2.mp4	Bajo	3689.95
Wood3.mp4	Medio	5180.78
Wood4.mp4	Medio alto	8090.87
Wood5.mp4	Alto	11952.7
Wood6.mp4	Muy alto	14928.66

6. Método de Evaluación

El método a utilizar en las pruebas es el denominado “ACR”, también conocido como “Índice por categoría absoluta”. En esta sección se presenta una descripción de dicho método, extraída de la recomendación ITU-T P.910 [3].

Descripción del método ACR

El método de los índices por categorías absolutas es un juicio de categorías en el que las secuencias de prueba se presentan una por vez y se califican independientemente en una escala de categorías. (Este método se denomina también método de evaluación con un solo estímulo.) El método especifica que después de cada presentación se invite a los sujetos a evaluar la calidad de la secuencia mostrada. En la Figura 1 se ilustra el diagrama de tiempos de la presentación del estímulo. Si se utiliza un tiempo de votación constante entonces el tiempo de votación debe ser igual o inferior a 10 s.

En la Figura 4 (extraída de [3]) se ilustra el diagrama de tiempos de la presentación del estímulo.

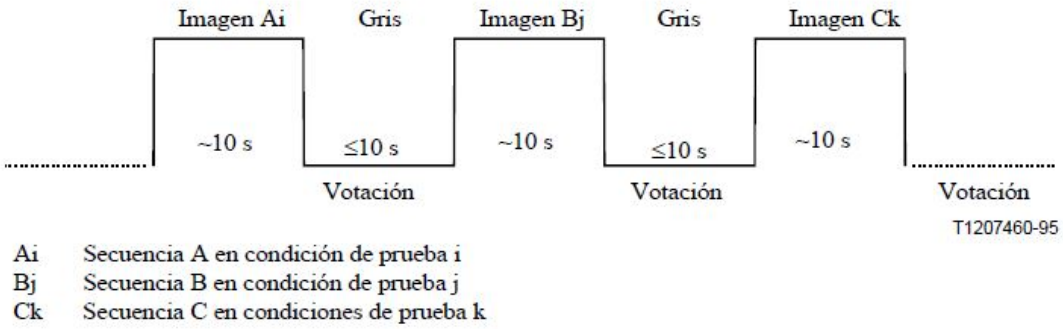


Figura 4
Presentación del estímulo en el método ACR

El tiempo de votación deberá ser igual o inferior a 10 s. El tiempo de presentación puede reducirse o aumentarse en función del contenido del material de prueba. Para evaluar la calidad global se debe utilizar la siguiente escala de cinco niveles, como se muestra en la tabla 8:

Cuadro 8
Puntajes y Niveles

Puntuación	Nivel
5	Excelente
4	Buena
3	Aceptable
2	Mediocre
1	Mala

7. Análisis de los resultados

7.1. Cálculo de MOS

El primer paso al analizar las evaluaciones subjetivas de calidad consiste en hallar el valor de MOS final correspondiente a cada contenido de video. Para cada contenido j se define $\mathbf{MOS}_j$ como:

$$\mathbf{MOS}_j = \overline{u}_j = \frac{1}{N}\sum_{1}^{N} u_{ij} \tag{1}$$

donde u_{ij} representa la evaluación del observador i para el contenido j y N es el número de observadores válidos.

En nuestro caso, se calculan valores de MOS promediando las opiniones de distintos observadores ubicados en la misma posición. En principio, no se calcula el MOS como promedio de opiniones de observadores en distintas posiciones.

7.2. Validación de calificaciones individuales

Se establece un criterio para validar las calificaciones individuales. Para ello se calcula la correlación entre las calificaciones individuales y las calificaciones promedio de todos los evaluadores por posición. Esta correlación se define según la ecuación (2) :

$$r_1(i) = \frac{\sum_{j=1}^{K} \overline{u}_j u_{ij} - \frac{\left(\sum_{j=1}^{K} \overline{u}_j\right)\left(\sum_{j=1}^{K} u_{ij}\right)}{K}}{\sqrt{\left(\sum_{j=1}^{K} \overline{u}_j^2 - \frac{\left(\sum_{j=1}^{K} \overline{u}_j\right)^2}{K}\right)\left(\sum_{j=1}^{K} u_{ij}^2 - \frac{\left(\sum_{j=1}^{K} u_{ij}\right)^2}{K}\right)}} \tag{2}$$

Donde:

- j es el índice de PVS
- $\overline{u}_j$ es el MOS calculado con todos los evaluadores por PVS j
- u_{ij} es la calificación individual del evaluador i para ese PVS j
- K es el número de PVSs

Para cada posición se calcula $r_1(i)$ para cada observador i, y se excluye a los que tienen $r_1(i)$ menor a 0,75. En caso de excluirse observadores, se vuelve a calcular el MOS de (1). Este proceso se realiza una única vez.

7.3. Cálculo del intervalo de confianza

Cuando se presenten los resultados de una prueba, todas las notas medias (MOS) deberán tener un intervalo de confianza asociado, que se obtiene a partir de la desviación típica y el tamaño de cada muestra. Se utiliza en este caso un intervalo de confianza del 95 %, que viene dado por las ecuaciones (3) , (4) y (5):

$$\mathbf{IC} = [\overline{u}_j - \delta_j, \overline{u}_j + \delta_j] \tag{3}$$

donde:

$$\delta_j = 1{,}96 \cdot \frac{S_j}{\sqrt{N}} \tag{4}$$

y

$$S_j = \sqrt{\sum_{i=1}^{N} \frac{(\overline{u}_j - u_{ij})^2}{N-1}} \tag{5}$$

donde:

- j es el índice de la PVS
- i es el índice del observador
- N es la cantidad de observadores
- $\overline{u}_j$ es el $\mathbf{MOS}_j$
- u_{ij} es la calificación individual del observador i para la PVS j.

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 %, siempre que la distribución de las notas individuales cumpla ciertos requisitos.

Calculando los intervalos de confianza para cada PVS en cada asiento, podremos tener una idea de cuan similar es la percepción de calidad del usuario que se encuentra en una posición, a la de un usuario que se encuentra en otra.

7.4. Test t de Student

Un análisis más riguroso que el presentado en la sección anterior, es realizar un test t de Student de dos muestras y varianzas desiguales, usando una distribución de dos colas para determinar si en efecto las calidades subjetivas dadas por los valores promedio de las muestras de un par de asientos no son iguales. La hipótesis nula H_0 en este caso sería que los observadores en diferentes asientos perciben la misma calidad para una PVS, y la hipótesis alternativa H_a es que los observadores en distintos asientos no perciben la misma calidad para una PVS. Para comparar las medias de dos poblaciones, puede usarse la estadística t, la cual se expresa como:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \tag{6}$$

donde $\overline{X}_i$, s_i^2, n_i denotan la media de la muestra, la varianza de la muestra, el tamaño de la i-ésima muestra y $i \in$ 1,2.

Se calculan $\overline{X}_i$ y s_i^2 como:

$$\overline{X}_i = \frac{1}{n_i} \sum_{k=1}^{k=n_i} X_k \tag{7}$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{k=n_i} (X_k - \overline{X}_i)^2 \tag{8}$$

donde X_k es la k-ésima opinión en un asiento para una PVS, e $i \in$ 1,2.

Al calcular la estadística t de esta forma y aproximándola con una distribución t de Student cuyo grado de libertad DF se define así

$$DF = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} \tag{9}$$

se puede calcular un valor de probabilidad p a partir de la estadística t que indica el grado al cual las medias de las dos poblaciones se consideran diferentes. Cuanto más pequeño es el valor p, más significativa es la diferencia entre las distribuciones de las poblaciones.

Un valor p menor a 0,05 indica una probabilidad muy baja de cometer un error tipo I (esto es, rechazar la hipótesis nula cuando es cierta). En tal caso, la hipótesis nula puede ser rechazada con seguridad, y puede concluirse que hay significado estadístico en que las dos ubicaciones perciben calidades diferentes. Un valor p mayor o igual a 0,05 significa que la hipótesis nula no puede ser rechazada con confianza. Sin embargo,todavía existe la posibilidad de cometer un error tipo II (esto es, no poder rechazar la hipótesis nula cuando de hecho la hipótesis alternativa es cierta).

El test será aplicado a todas las degradaciones de cada uno de los videos, tomando pares de ubicaciones. Cada ubicación será comparada con la ubicación central, ubicación 2 en la figura 1. Con este criterio se obtendrán 504 resultados del test.

8. Referencias

[1] ITU-R BT.500-13, “Methodology for the subjective assessment of the quality of television pictures, ” International Telecommunication Union, 2012.

[2] https://github.com/Telecommunication-Telemedia-Assessment/SITI

[3] ITU-T, “Subjective video quality assessment methods for multimedia applications”, Recommendation ITU-R P 910, Sep 1999.

[4] E. G. Richardson, Iain (2003). H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia. Chichester: John Wiley and Sons Ltd.

[5] https://www.ffmpeg.org/

[6] Sitio web de Avisynth: http://avisynth.nl

[7] Sitio web de MVTools https://avisynth.org.ru/mvtools/mvtools2.html

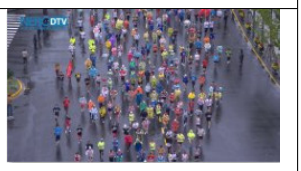
[8] Sitio Web de VirtualDub: http://www.virtualdub.org/

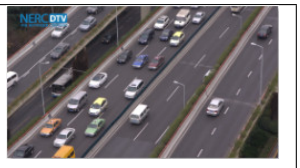
[9] *Shanghai Jiao Tong University.* [Online]. Available: http://medialab.sjtu.edu.cn/web4k/index.html

[10] *Ultra Video Group.* [Online]. Available: http://ultravideo.cs.tut.fi/

A. Anexo 1 - Secuencias de video seleccionadas

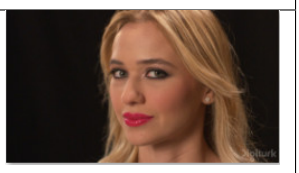
Las secuencias de video utilizadas fueron descargadas del sitio de *Shanghai Jiao Tong University* [9]

Título	Descripción	Primer Cuadro
Bund Nightscape	Plano aéreo de una ciudad en cámara rápida, durante la noche.	
Campfire Party	Muestra llamas ante el equipo de NERC-DTV en una fiesta con una fogata.	
Construction Field	Muestra una excavadora en un sitio de construcción.	
Fountains	Muestra chorros verticales de una fuente de agua frente a un edificio alto	
Marathon	Muestra la escena de las primeras etapas de la 2012 Shanghai International Marathon Race	
Runners	Muestra muchos corredores en medio de la 2012 Shanghai International Marathon Race	
Rush Hour	Muestra muchos estudiantes yendo a la cantina o al dormitorio después de clases.	
Scarf	Muestra unas bufandas moviéndose al viento.	
Traffic and Building	Muestra tráfico con una ciudad de fondo	

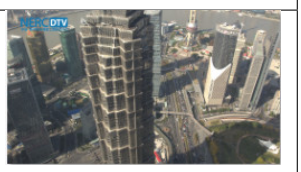
Traffic Flow	Muestra caminos con automóviles moviéndose en distintas direcciones	
Tree Shade	Muestra un lugar con mucha vegetación y un árbol cuyas hojas se mueven al viento.	
Wood	Muestra un bosque en el campus de SJTU con rayos del sol penetrándolo.	
Mobile	Muestra un tren de juguete circulando y otros objetos en movimiento.	
Coastguard	Muestra un barco navegando bajo el sol.	

B. Anexo 2 - Secuencias de video descartadas

Las siguientes secuencias de video descartadas fueron descargadas del sitio de *Ultra Video Group* [10].

Título	Descripción	Primer Cuadro
Beauty	Muestra la cara de una mujer que pestañea y cuyo pelo se mueve con el aire.	
Bosphorous	Muestra un bote navegando en un río, con un puente y una ciudad de fondo.	
HoneyBee	Muestra a una abeja moviéndose entre distintas flores.	
Jockey	Muestra a un jinete cabalgando en una carrera.	
ReadySetGo	Muestra el comienzo de una carrera de caballos.	
ShakeNDry	Muestra a un perro sacudiéndose el agua para secarse.	
Yacht Ride	Muestra una pareja en un yate navegando.	

Las siguientes secuencias de video descartadas fueron descargadas del sitio de *Shanghai Jiao Tong University* [9]

Library	Muestra el exterior de una biblioteca con gente pasando.	
Residential Building	Muestra el exterior de un edificio residencial.	
Tall Buildings	Muestra edificios altos en Lujiazui, Pudong New District Shanghai	

C. Anexo 3 - Instrucciones para los observadores

Al inicio de la prueba se les explica a los observadores la forma en que se realizará la misma, así como también los detalles para votar.

1. Se indica a que red de wifi deben conectar sus smartphones y las credenciales de acceso. En caso de que alguien no tenga un celular con navegador, se le proveerá uno.
2. Se exhorta a los participantes que utilizan celular propio a silenciar sus celulares durante la prueba y a prestar atención solamente a la prueba (no a mensajes entrantes, etc.).
3. Se indica a cada participante cuál es su número de asiento.
4. Se les explica como registrarse a la aplicación que recoge los datos y los votos. Para esto se provee la IP y se les comenta que deben completar los datos que se piden.
5. Se explica sucintamente como funciona el programa en cuanto a la reproducción de los videos y como transcurrirá la votación.
6. Se aclara que la votación será exclusivamente sobre la calidad de imagen del video y no sobre el contenido.

La exposición oral inicial se redacta a continuación, para que la prueba sea la misma, sin importar que cambie la persona que da las indicaciones.

Estimados participantes:

Sean bienvenidos a la prueba de evaluación de calidad de video, en el marco de nuestro proyecto de fin de carrera en la Facultad de Ingeniería de la Universidad de la República. Desde ya agradecemos su participación.

Si alguien necesita salir de la sala (por ejemplo, para ir al baño), que lo haga ahora. Una vez comenzada la prueba no se podrá abandonar su asiento hasta que la misma finalice.

Pausa para ir al baño.

Pasamos a comentarles cómo se desarrollará la prueba. Se mostrará una serie de 84 videos sin sonido de 12 segundos de duración cada uno. Después de cada video, el mismo deberá ser calificado en una escala de 5 posibles categorías:

- Excelente
- Bueno
- Aceptable
- Mediocre
- Malo

Solo se podrá calificar una vez cada video. Los mismos contenidos se verán varias veces, con diferentes tipos de degradaciones o problemas de calidad. Por favor, realice cada calificación en forma independiente de las anteriores.

Desde el navegador web de los celulares que les vamos a proveer a cada uno, completen los datos pedidos por la aplicación. El número de cédula tendrá que ingresarse sin puntos y sin el guion. El número de asiento que les pedirá la aplicación, es el que está pegado a cada una de sus sillas.

Pausa. Se le entregan los celulares a todos los usuarios.

Esta misma aplicación es la que usarán para emitir sus votos luego de cada video. En el televisor se les indicará cuándo comienza el siguiente video, y cuando el video finaliza les pedirá que lo califiquen. En ese momento aparecerán en sus celulares las 5 categorías para que califiquen. No es necesario apurarse a votar, ya que la aplicación espera a recibir todos los votos antes de pasar al siguiente video.

Pausa. También se supervisa y verifica que los usuarios completen satisfactoriamente esta etapa.

Les vamos a solicitar encarecidamente que silencien sus celulares. A aquellos que estén usando sus propios celulares para la prueba, por favor, rogamos se abstengan de usarlos con otros propósitos que no sea el de votar.

Haremos especial hincapié, en que se deberá evaluar únicamente la calidad de cada video, independientemente de su contenido, el cual puede llamar más o menos nuestra atención dependiendo de nuestros gustos.

¿Alguna pregunta?

Muchas gracias.