UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

DETECCIÓN AUTOMÁTICA DE PUBLICIDAD EN SEGMENTOS DE VIDEO

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN

CRISTIAN RODRIGO VÁSQUEZ LEAN

PROFESOR GUÍA:
JUAN BARRIOS NÚÑEZ

MIEMBROS DE LA COMISIÓN:
BENJAMÍN BUSTOS CÁRDENAS
NANCY HITSCHFELD KAHLER
RODRIGO SALAS FUENTES

SANTIAGO DE CHILE
2020

Resumen

La presente tesis tiene como objetivo entregar una solución tecnológica que permita soportar la detección automática de publicidad en segmentos de video. Esta solución permitirá cuantificar de forma automatizada la participación de las compañías en términos de emisión de comerciales y entender cuál es el comportamiento del mercado publicitario.

Este trabajo trata el problema que afecta a compañías que publicitan comerciales en la televisión y no poseen una herramienta para saber de manera oportuna y precisa los comerciales que están apareciendo en la televisión en cada momento. La necesidad de esta información es relevante para la toma de decisiones exitosas, en un mercado cada vez más competitivo.

Se desarrolló una solución que permite detectar, mediante técnicas de Recuperación de Información Multimedia, la aparición de comerciales en las diferentes programaciones diarias de los canales de televisión. Cada parrilla programática diaria fue descargada en formato de video, desde los servicios de *streaming online* que provee cada canal. Desde cada video se extrajeron características de imagen y audio, con el propósito de generar descriptores de contenidos que permiten lograr una detección de comerciales efectiva y eficiente.

Los distintos tipos de descriptores fueron evaluados utilizando métricas de *Precision*, *Recall* y *Mean Average Precision*, con el propósito de identificar los descriptores de mejor desempeño. Las evaluaciones arrojaron los mejores resultados al combinar descriptores de imagen y audio, mientras que al utilizar descriptores por separado, los resultados disminuyeron drásticamente en cantidad de recuperaciones, a pesar de identificar correctamente el comercial.

Por otro lado, al analizar los datos obtenidos, se generaron estadísticas y diferentes visualizaciones que permitieron reconocer diferentes comportamientos sobre la publicidad emitida en la televisión local, para un mes en particular en distintos momentos del día y la semana. Al identificar las emisiones de comerciales que se realizan en un determinado período de tiempo, es posible descubrir y analizar la estrategia publicitaria que las compañías emplean en la televisión.

Dedico mi trabajo a Evelyn,
quien estuvo a mi lado en este bonito desafío

Tabla de contenido

Índice de tablas

Índice de ilustraciones

1. Introducción

El contenido multimedia, como una imagen, una pista de audio o video, corresponde a una de las definiciones con mayor impacto en la última década. No solo es utilizado por los teléfonos móviles al intercambiar datos en un chat, o al visitar una red social, también se encuentra en la mayoría de los dispositivos electrónicos que las personas utilizan diariamente [R1].

Poco a poco ha ido sumando mayor protagonismo en el entorno que rodea a las personas, producto que las compañías se dieron cuenta que es un recurso que logra capturar por mayor tiempo su atención, abriendo nuevas posibilidades para atraer potenciales clientes a su negocio. Esto se produce porque el cerebro humano procesa la información visual más rápidamente, en comparación a textos con información.

Actualmente las compañías utilizan el contenido multimedia para crear y emitir anuncios publicitarios con diferentes propósitos, como anunciar hitos importantes, vender productos, o bien, ofrecer servicios.

1.1. Contexto

En la televisión todos los días se busca captar la atención de los televidentes con diferentes anuncios, promociones, ofertas, lanzamientos de productos, entre otros.

Un comercial corresponde a un anuncio audiovisual que busca enviar un mensaje particular, con el objetivo de captar la atención de una audiencia específica, como la televisiva, y cuya duración por lo general no sobrepasa los 60 segundos [R2]. Cada comercial tiene un tiempo de duración determinado y puede llegar a ser emitido varias veces al día.

La televisión se mantiene como uno de los medios electrónicos más importantes, con un promedio de 170 minutos diarios en tiempo de visualización [R3], permitiéndole alcanzar diariamente cerca del 70% de la población de un país, el 90% en una semana y cerca del 100% al mes [R4].

En la actualidad, este medio de comunicación resulta de mucha importancia para las compañías que ofrecen productos o servicios. Esto se debe a una fuerte correlación entre la inversión de dinero en anuncios publicitarios y el rápido aumento de los ingresos, producto del alcance de la población que puede cubrir.

Un ejemplo sobre ello corresponde al evento deportivo de *Super Bowl* que se realiza en los Estados Unidos, marcando *peaks* de audiencia televisivos en todo el mundo, y en el cual las compañías invierten enormes cantidades de dinero para presentar nuevos productos o informar sobre un hito importante cada año [R5].

El medio local tampoco se queda atrás. Desde el año 1978 se realiza el evento solidario llamado Teletón, cuyo objetivo busca recolectar dinero para financiar la rehabilitación de discapacitados. Este financiamiento se basa en un 50%, gracias al aporte de las empresas a la fundación. Mediante campañas publicitarias, la televisión y otros medios de comunicación buscan persuadir a las personas para comprar productos asociados al evento, con el objetivo de apoyar a la fundación para ampliar su dotación de personal, rehabilitar a más personas, o crear nuevos centros de rehabilitación.

Si bien la televisión ha sido un nicho importante para el mercado publicitario, la aparición de medios digitales como web de noticiarios, redes sociales, entre otros, ha motivado a las compañías para marcar presencia en cada uno de estos medios. Su objetivo es capturar una nueva audiencia de potenciales consumidores para explotar el mercado publicitario audiovisual.

La mayoría de los portales web de los canales de televisión, nacionales e internacionales, dispone de servicios de libre acceso a información relacionada a la programación diaria, como también, *streamings online* de televisión que permiten capturar y reproducir el contenido audiovisual que está siendo emitido por la señal televisiva.

Por lo anterior, la participación en medios digitales ha producido que la inversión publicitaria aumente. Por ejemplo, durante el año 2017 el gasto en publicidad digital alcanzó $209 billones de dólares, ocupando el 41% del mercado publicitario. Por otra parte, la publicidad televisiva gastó $178 billones de dólares en el mismo año, ocupando el 35% del mercado publicitario [R6]. Se espera que para el año 2020 la publicidad digital alcance el 50% del gasto en el mercado publicitario.

Para concretar la inversión publicitaria y participar en los medios de comunicación, las compañías delegan a sus *Product Managers* la responsabilidad de posicionar marcas. Mediante la utilización de la información proporcionada por agencias publicitarias, los *Product Managers* conectan con diferentes entidades para realizar patrocinios, publicitar productos y competir en el mercado.

Dado este contexto, la Ciencia de los Datos ofrece una visión diferente a la información que las agencias publicitarias y *Product Managers* actualmente utilizan. Por ejemplo,

permitiendo monitorear automáticamente la televisión, analizar datos en línea, interactuar con datos recuperados directamente desde la televisión, etc.

1.2. Problema

Debido a la gran cantidad de información publicitaria que se genera día a día en la televisión, descubrir la estrategia publicitaria de una compañía podría ser posible si se analizan sus datos oportunamente en el tiempo.

La información publicitaria corresponde a contenido audiovisual, compuesto por imágenes y audio, que es emitido en la televisión en diferentes horarios del día.

Si bien en el medio existen empresas dedicadas al mercado de investigación de medios de comunicación, profesionales dedicados a cubrir tareas publicitarias, como los *Product Managers*, no disponen de una herramienta tecnológica que se ajuste a sus necesidades para capturar y analizar esta información.

Parte de las necesidades están relacionadas a entender el comportamiento del mercado publicitario, observar la competencia y reaccionar oportunamente de acuerdo a los cambios que ocurran.

Los *Product Managers* no solo interactúan en el medio local. Generalmente están ubicados estratégicamente por su compañía a nivel regional, posicionando marcas a nivel latinoamericano, o bien, abarcando varios países.

Para diseñar una buena estrategia publicitaria, no solo se debe considerar información sobre la publicidad que actualmente es emitida en la televisión local. También resulta interesante entender cómo es el comportamiento de la competencia a nivel internacional.

Una de las principales fuentes de información que utilizan los *Product Managers* corresponde a la proporcionada por agencias publicitarias, cuyo contenido corresponde a planillas que solo registran información actual del mercado publicitario, con escasos atributos útiles para responder diferentes preguntas que pueden surgir al realizar algún estudio.

Si un *Product Manager* debe posicionar una marca en diferentes países, puede llegar a depender de más de una agencia publicitaria, disminuyendo las posibilidades de obtener la información necesaria en tiempo y forma.

Con el objetivo de cubrir el problema expuesto de forma transversal, se dará respuesta a las siguientes dos hipótesis utilizando técnicas de Recuperación de Información Multimedia:

- Es posible identificar apariciones de comerciales buscando secuencias duplicadas en videos utilizando contenido visual
- Es posible mejorar el resultado al combinar el contenido de audio y video, comparado con el contenido visual

1.3. Situación Actual

En la actualidad, es importante considerar que el estado del arte está cubierto por diferentes soluciones cuyo objetivo busca dar respuesta a la necesidad expuesta como problema. Algunas de ellas son:

- Comskip [R12]
- Showanalyzer [R13]

Estas soluciones se rigen por parámetros, que requieren de esfuerzo en su configuración y en general proporcionan resultados de baja calidad. Además, se basan en condiciones especiales para emitir comerciales en televisión, tales como anteponer y finalizar cada comercial con un cuadro *(frame)* de color negro, o poseer una duración de 30 segundos para realizar una detección correctamente [R7].

En Chile, este tipo de soluciones no lograrán una correcta detección de comerciales en la televisión, debido a que poseen un sesgo producido por las técnicas utilizadas. A diferencia de otros países, los comerciales en Chile no se rigen por un estándar, como anteponer y finalizar los comerciales con un *frame* de color negro.

En resumen, cada comercial puede registrar un comportamiento diferente cuando es emitido, convirtiendo esta solución en una opción poco flexible para trabajar.

Para cubrir la necesidad que los *Product Managers* presentan, se requiere mejorar las técnicas que proveen las soluciones mencionadas anteriormente.

1.4. Solución

Teniendo en cuenta la problemática señalada, este proyecto de tesis busca desarrollar un **detector automático de publicidad televisiva**, permitiendo consultar sobre datos históricos, generar estadísticas, reportes y *dashboards*.

Para empezar, la captura de videos de televisión se realizará desde el contenido que proporcionan los medios locales, tales como: servicios de *streaming*, videos subidos a los portales de la web, redes sociales.

Mediante la búsqueda, identificación y comparación de información audiovisual utilizando técnicas de Recuperación de Información Multimedia y Ciencia de Datos, se construirá una base de datos con las ocurrencias detectadas para cada comercial.

Para materializar esta herramienta se debe disponer de una infraestructura de extracción, procesamiento, transformación y consulta de datos multimedia, en ejecución y disponible para su acceso continuo.

En términos técnicos, lo anterior se traduce en el desarrollo de una herramienta que permita a los *Product Managers* reaccionar con una estrategia publicitaria oportuna frente a la competencia y las necesidades que demanda el mercado publicitario, como también, entender el comportamiento que ha tenido su negocio a lo largo del tiempo.

Por último, esta herramienta ampliará el conocimiento sobre el comportamiento del mercado publicitario, complementando el contenido proporcionado por las agencias publicitarias con información relevante, pertinente y a la cual podrán acceder sin restricciones.

1.5. Objetivos

El objetivo general de este proyecto de tesis consiste en la construcción de una solución que permita detectar automáticamente apariciones de publicidad en televisión y la generación de estadísticas sobre las detecciones registradas.

Los objetivos específicos del proyecto son los siguientes:

- Implementar un módulo para descargar televisión online
- Implementar un algoritmo de detección de comerciales basado en características visuales, de audio y audiovisuales
- Implementar algoritmos para calcular estadísticas de aparición de comerciales

- Evaluar la calidad de detección de comerciales utilizando un set de datos de pruebas
- Comparar la calidad de detección al utilizar características visuales, de audio y audiovisuales

Con la finalidad de desarrollar la solución, realizar pruebas y evaluar sus resultados, se ha creado un *dataset* de pruebas cuyo contenido corresponde a un mes de programación televisiva etiquetada del canal CHV, entre los días 03 de Septiembre y 01 de Octubre del año 2018.

El set de datos suma un total 485 horas de video, 643 comerciales distintos y 7492 emisiones agrupadas en 16 etiquetas que registran 43 horas, 55 minutos y 12 segundos en tiempo de reproducción. Las etiquetas tienen como propósito categorizar, de forma general, el contenido que registra cada comercial.

En total, el contenido multimedia registra 210 GB almacenados en formato mp4, una resolución de 768x432 píxeles, calidad a color y 29.97 fps.

2. Marco Teórico

Los principales desafíos para este trabajo estarán relacionados a la extracción de características, la generación de descriptores, el tiempo de procesamiento para identificar y obtener resultados, y realizar un estudio para identificar cual es la mejor configuración para los diferentes tipos de contenido multimedia que se utilicen.

Para extraer características desde el contenido multimedia, como las imágenes, librerías de visión computacional, como OpenCV [R15], son esenciales para generar los descriptores visuales que permiten realizar una tarea de búsqueda por imagen.

El audio, por otro lado, requiere de otro tipo de librerías para poder acceder a su contenido. Librerías como LibROSA [R16], permiten descomponer el audio y realizar diferentes análisis, como también, extraer características para generar descriptores que permiten realizar una tarea de búsqueda por audio.

Con la finalidad de procesar una consulta y calcular resultados, se investigarán alternativas relacionadas a las búsquedas por similitud [4, P15] con el propósito de identificar la que mejor rendimiento y resultados entregue en esta tarea.

Finalmente, los estudios sobre los resultados de Recuperación de Información se realizarán utilizando métricas de *Precision, Recall y Mean Average Precision,* con el propósito de identificar cual es la mejor configuración para realizar una detección.

2.1. Recuperación de Información Multimedia *(Multimedia Information Retrieval)*

La Recuperación de Información estudia cómo representar, organizar, almacenar y acceder a información existente en documentos. Su objetivo más común es localizar dentro de un conjunto los documentos que son relevantes a una necesidad de información del usuario. Principalmente ha estado relacionada a textos como fuentes de datos.

Para recuperar datos estructurados desde una base de datos relacional, la tarea es bastante sencilla de ejecutar. Sin embargo, cuando los atributos de los datos corresponden a contenido multimedia, la recuperación de datos estructurados ya no es una alternativa viable de ejecución.

Desde los años 90', la difusión de contenido multimedia, como imágenes, música, y videos, dio inicio a nuevas ramas de estudio que buscan desarrollar técnicas para poder recuperar estos objetos multimedia.

La Recuperación de Información Multimedia tiene por objetivo procesar y localizar dentro de un conjunto de documentos multimedia (audio, imagen, video, objetos 3d, etc.) los que son relevantes a la necesidad de información del usuario.

Esta rama de estudio busca recuperar información sobre los objetos multimedia, cuyo contenido puede ser utilizado para analizar y procesar datos, superando los resultados obtenidos en la Recuperación de Información tradicional [2].

2.2. Visión Computacional *(Computer Vision)*

La Visión Computacional estudia métodos para adquirir, procesar, y analizar imágenes del mundo real con el fin de que sean utilizadas por un computador.

Los seres humanos capturan las estructuras que los rodean. Esa información es enviada al cerebro y puede ser consultada para recordar, describir e incluso manipular su contenido para el propósito que se estime necesario.

Si una persona intenta crear un dibujo con esta información, con mucha dificultad y tiempo podría producir una representación que se parezca a lo que su cerebro registra en

memoria. Es decir, la percepción visual que los humanos tienen sobre la realidad puede resultar muy diferente a lo que realmente es.

En términos computacionales, recuperar la percepción visual no es una tarea sencilla y requiere incorporar diferentes métodos para comprender y simular las capacidades que tiene el cerebro humano.

La disciplina de Visión Computacional busca interpretar, mediante un computador, lo que el cerebro humano es capaz de percibir del mundo que visualiza, analiza imágenes para obtener propiedades de los objetos como su forma, ubicación, iluminación y distribución de colores.

Una imagen corresponde a una señal de 2 dimensiones (ancho y largo) que posee canales (1, 3 o 4) y usualmente una profundidad de 8 bits.

Un píxel *(picture element)* corresponde a un punto de la imagen, cuyo valor de intensidad varía entre 0 y 255. El valor "0" corresponde al color negro absoluto, mientras que "255" al color blanco absoluto.

La Imagen 1 muestra un mapa con una ampliación en una porción de la imagen, con el propósito de visualizar los valores que se registran por cada píxel.

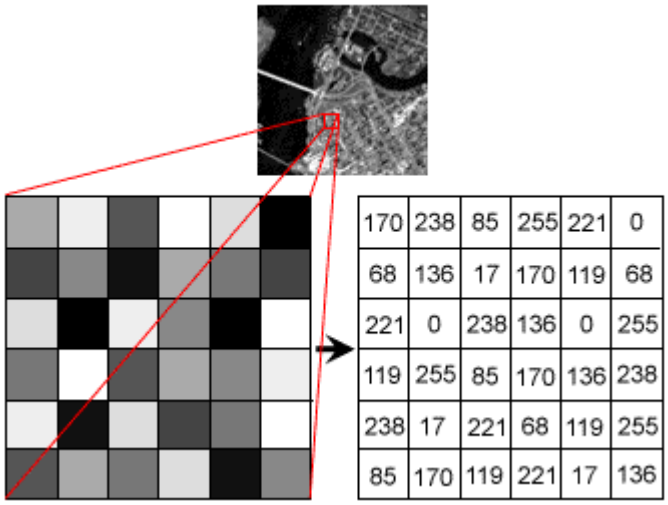

Imagen 1 - Visualización de píxeles en imagen [IMG 1]

En la Imagen 2, cada píxel tiene una ubicación que se determina mediante la fila y columna en la cual esté posicionado, tomando como punto base (0,0) la esquina superior izquierda.

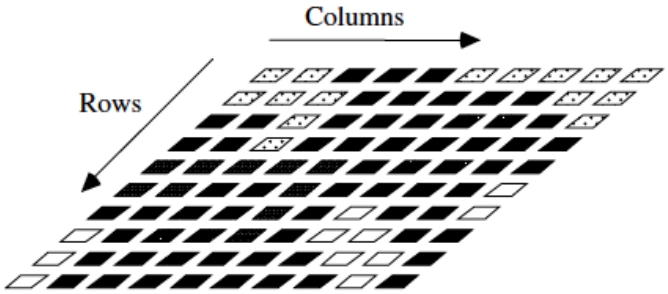

Imagen 2 - Distribución de píxeles [IMG 2]

Una vez extraídos los datos de las imágenes, pueden ser procesados para aplicar operaciones, tales como:

- punto a punto (por píxel)
- lineales (convoluciones)
- no lineales

A continuación, las Imágenes 3 y 4 muestran cómo se realiza una operación de convolución al multiplicar la imagen por un *kernel* y cuál es el resultado que se obtiene luego de realizar la operación.

45	60	98	127	132	133	137	133
46	65	98	123	126	128	131	133
47	65	96	115	119	123	135	137
47	63	91	107	113	122	138	134
50	59	80	97	110	123	133	134
49	53	68	83	97	113	128	133
50	50	58	70	84	102	116	126
50	50	52	58	69	86	101	120

*

0.1	0.1	0.1
0.1	0.2	0.1
0.1	0.1	0.1

=

69	95	116	125	129	132
68	92	110	120	126	132
66	86	104	114	124	132
62	78	94	108	120	129
57	69	83	98	112	124
53	60	71	85	100	114

Imagen 3 - Operación lineal en una imagen [IMG 3]

Imagen 4 - Izquierda: Imagen en escala de grises;
Derecha: Imagen después de una convolución [IMG 4]

Actualmente, la disciplina de Visión Computacional es utilizada en una gran variedad de aplicaciones pertenecientes al mundo real, dentro de las cuales destacan: reconocimiento óptico de caracteres (OCR), vigilancia, biometría, captura de movimiento, reconocimiento de imágenes, entre otros [3].

2.2.1. Procesamiento Digital de Imagen *(Digital Image Processing)*

El Procesamiento de Imágenes Digitales corresponde a un método que se refiere a las técnicas para mejorar la calidad o modificar la información en imágenes digitales.

Este método está relacionado al Procesamiento Digital de Imágenes por medio de un computador, donde la imagen posee un número finito de elementos con una ubicación particular [7].

Cuando “x” e “y” junto a la amplitud de los valores de “f” son todos finitos, se puede llamar a la imagen como “imagen digital”.

Para trabajar las imágenes es necesario extraer los elementos por imagen, denominados píxeles, los cuales aportan un valor de intensidad a la imagen y finalmente permiten entender lo que se visualiza.

Como uno de los objetivos de este proyecto de tesis es identificar imágenes que son similares, es necesario aplicar técnicas de extracción y transformación de datos para poder aplicar cálculos que permitan entender si una imagen es similar a otra.

El contenido multimedia usualmente utiliza imágenes a color, donde generalmente su representación está dada por el modelo de colores RGB *(red, green, blue)*. Al sumar estas 3 capas de colores, es posible obtener la imagen digital tal como la perciben los humanos (ver Imagen 5).

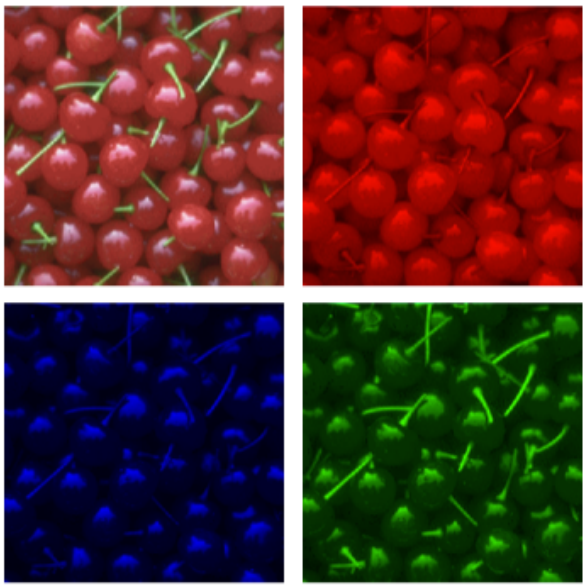

Imagen 5 - Canales RGB de una imagen [IMG 5]

Una buena práctica para manipular imágenes es convertir su contenido a escala de grises y disminuir sus dimensiones a un tamaño menor WxH. El propósito de lo anterior es representar la imagen en una baja resolución y disminuir la cantidad de capas (ver Imagen 6).

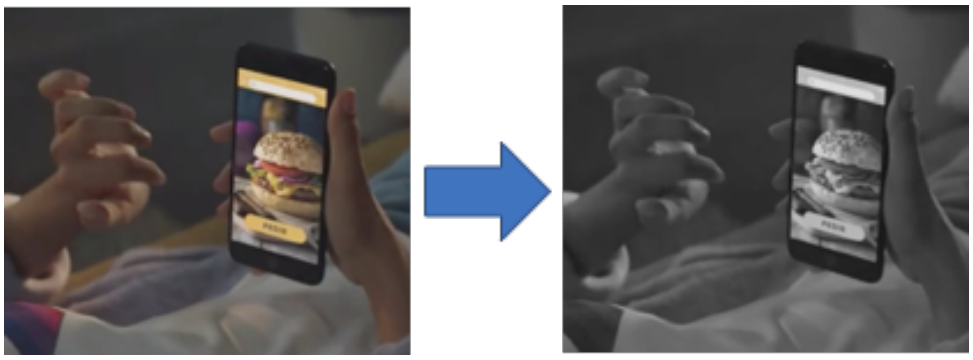

Imagen 6 - Transformación a escala de grises

2.3. Representación del Audio *(Audio Representation)*

El término audio está relacionado a la transmisión, recepción o reproducción de sonidos que yacen dentro de los límites del oído humano. Estos sonidos son transmitidos mediante una señal emitida por un objeto que vibra, cuya frecuencia es medida en Hertz (Hz) y desde la cual se obtienen muestreos *(sample rate)* por cierta cantidad de segundos.

Los muestreos para la voz usualmente son de [8 kHz], para *compact disc* entre [16 kHz - 44.1 KHz] y para los discos *blu-ray* entre [48 KHz - 192 KHz].

Las señales percibidas por el oído corresponden a ondas de aire que ingresan por el canal auditivo hasta el tímpano, cuya función se encarga de mover una serie de pequeños huesos en el oído medio.

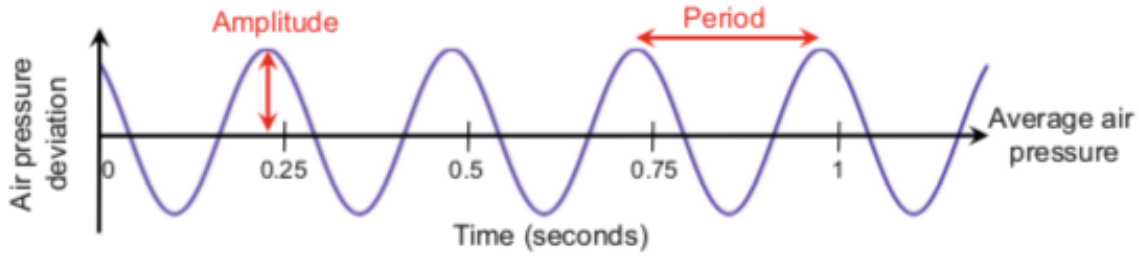

Imagen 7 - Representación de una señal [IMG 7]

El oído humano puede percibir entre los 20 Hz y 20000 Hz, siendo más sensible a frecuencias que fluctúan entre los 2000 Hz y 5000 Hz. Sin embargo, diferentes estudios indican que, con el paso de los años, el oído humano va perdiendo sensibilidad para las frecuencias más altas [R8].

Para tener una idea más clara respecto a lo que puede percibir el oído humano, se definieron propiedades fundamentales que tienen relación a la Representación del Audio. Entre ellas destacan: frecuencia, tono, dinámica, intensidad, ruido y timbre [6].

Lograr que un computador simule la capacidad que tiene el cerebro humano de interpretar las señales que recibe desde el oído, no es un trabajo sencillo. Al igual que la Visión Computacional, la Representación del Audio requiere incorporar diferentes métodos para comprender y simular las capacidades que tiene el cerebro humano.

En primer lugar, se requiere utilizar un micrófono para capturar las ondas de sonido y convertirlas en señales eléctricas, que luego, son codificadas en términos de amplitud cada cierto tiempo fijo *(Pulse Code Modulation, PCM)*.

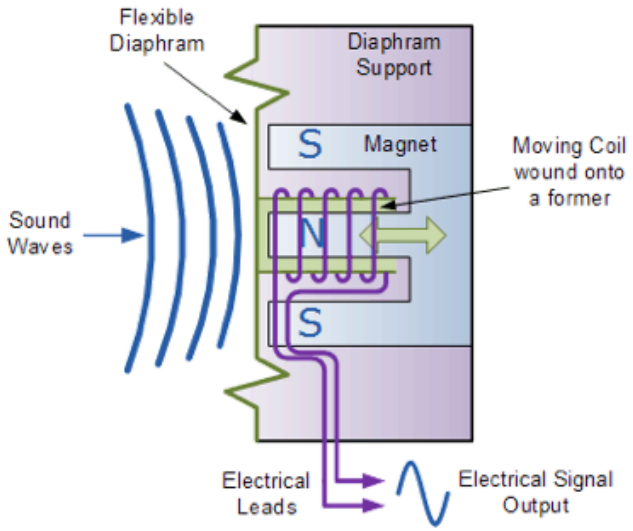

Imagen 8 - Funcionalidad de un micrófono

Cada valor de amplitud utiliza una resolución de 16, 24 o 32 bits. Es importante señalar que la codificación del audio puede incluir pequeñas desviaciones en el tono, timbre, entre otros.

Como se ha expuesto anteriormente, el sonido puede ser representado por una onda. Es decir, si los puntos altos y bajos de presión de aire se repiten de una forma alternada y regular, el resultado de la onda puede ser periódica. Mientras más alta sea la frecuencia de la onda, más alto será el sonido.

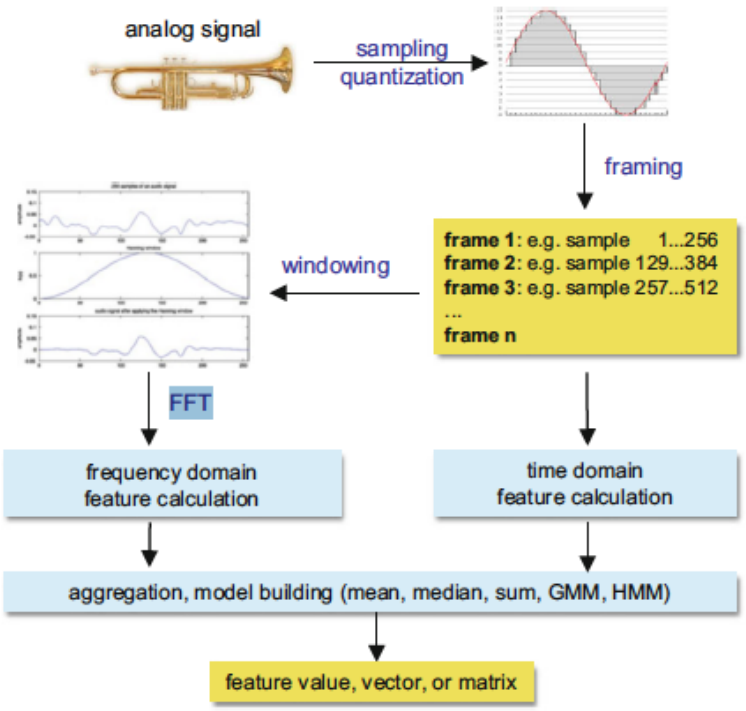

Imagen 9 - Flujo señal análoga y extracción de datos [IMG 9]

Al extraer un pequeño fragmento de tiempo de la onda de sonido, se puede visualizar con mayor detalle como varía la amplitud de la onda en el tiempo, tal como muestra la Imagen 10.

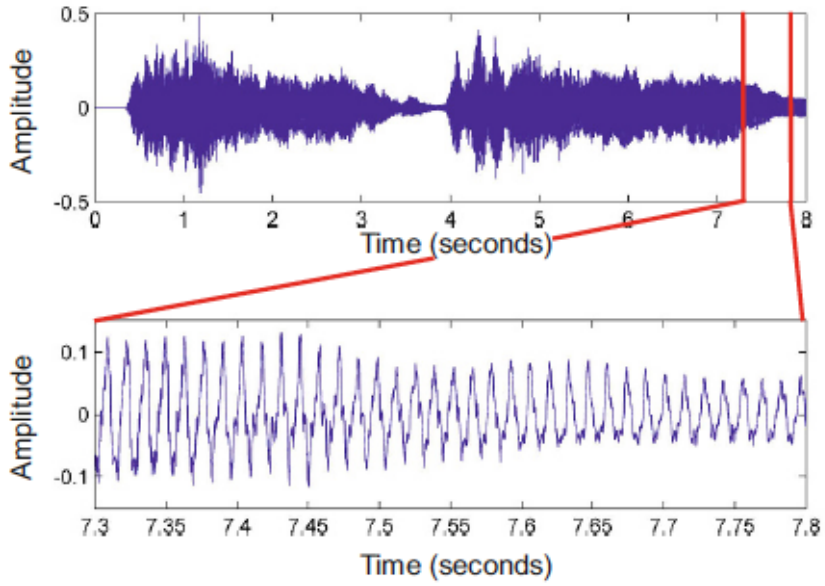

Imagen 10 - Representación de una pista de audio [IMG 10]

Se puede decir que existen tres niveles de abstracción para el audio, tal como muestra la Imagen 11. Alto, medio y bajo. Cada uno de ellos representa diferentes características que pueden ser utilizadas para diferentes propósitos.

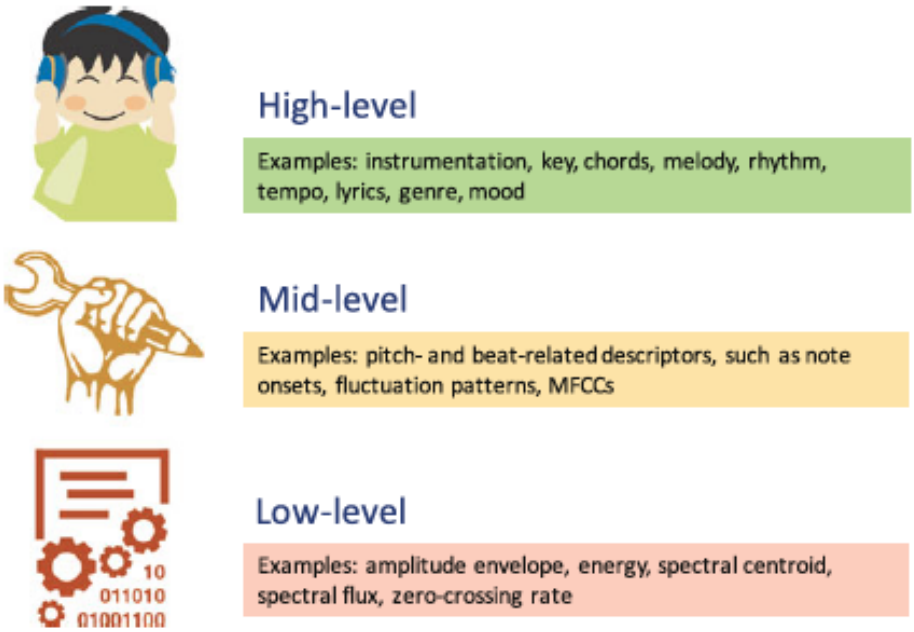

Imagen 11 - Niveles de abstracción del audio [IMG 11]

En la actualidad, la representación y procesamiento del audio es utilizado por diferentes aplicaciones, tales como: reconocimiento automático de canciones (shazam), identificación de estilos de música, identificación de comandos por voz, sincronización de audio, entre otros.

2.3.1. Transformada Rápida de Fourier *(Fast Fourier Transform, FFT)*

Corresponde a la transformación desde una señal del dominio del tiempo, a una del dominio de la frecuencia.

El teorema de Fourier, que subyace a la transformación respectiva, establece que cualquier función continua y periódica puede representarse como la suma de las ondas sinusoidales y coseno que oscilan a diferentes frecuencias, entregando como resultado un conjunto

discreto de valores complejos que describen el espectro de frecuencias de la señal de entrada.

2.3.2. Ventanas *(Windowing)*

Consiste en procesar una pista de audio para capturar pequeñas porciones de tiempo entre 10 a 100 ms, generando entre 256-8192 muestras (potencia 2).

Una vez obtenidas las muestras, se calcula la FFT con el propósito de obtener la energía de cada frecuencia.

Para evitar saltos, previo a la FFT se multiplica la ventana por una función que suavice bordes (función de Hann).

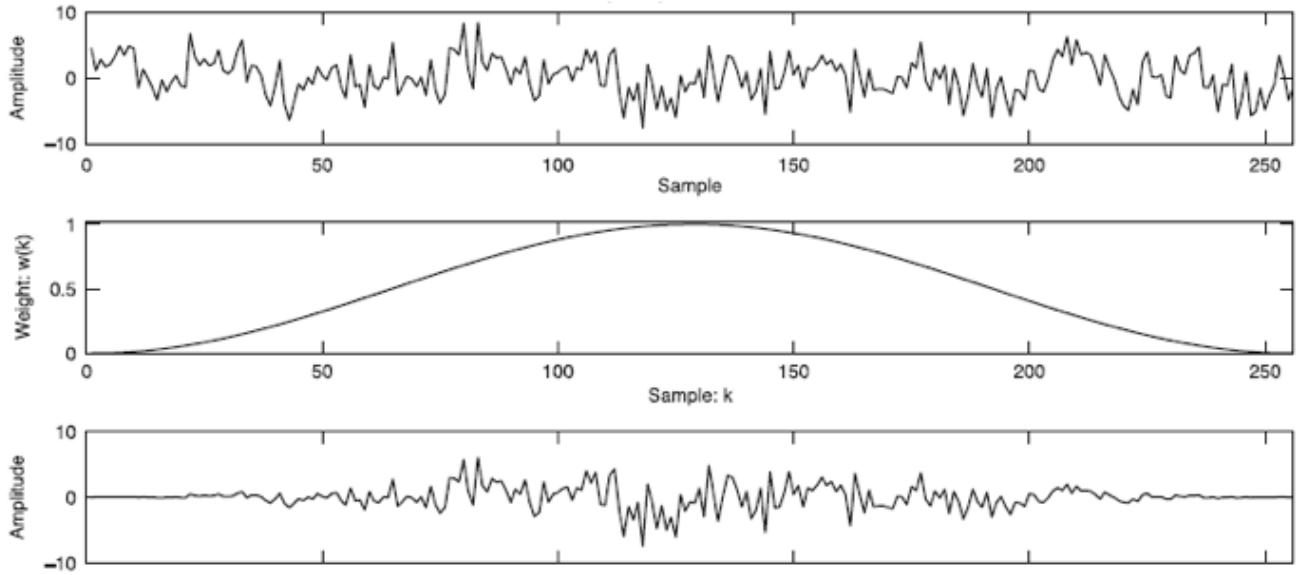

Imagen 12 - Windowing de 256 muestras [IMG 12]

La ventana de Hamming se encuentra con mayor frecuencia en el análisis de señales de audio para parámetros en el dominio de la frecuencia.

2.3.3. Escalas Mel *(Mel Scale)*

La escala Mel corresponde a una distorsión lineal gradual del espectro. Es decir, el oído humano interpreta los tonos *(pitch)* por sus frecuencias y no los puede interpretar de forma lineal.

Para compensar esto, luego de diferentes experimentos y estudios del tono en los años 40' se desarrolló la escala Mel, con el objetivo de entender el sistema auditivo en una escala lineal.

Dada las magnitudes sobre las frecuencias para cada ventana, la frecuencia primero es convertida a escala Mel calculando el logaritmo de las frecuencias.

Al aplicar la escala Mel, se puede obtener como resultado una descripción más precisa de la señal desde el punto de vista del sistema auditivo humano.

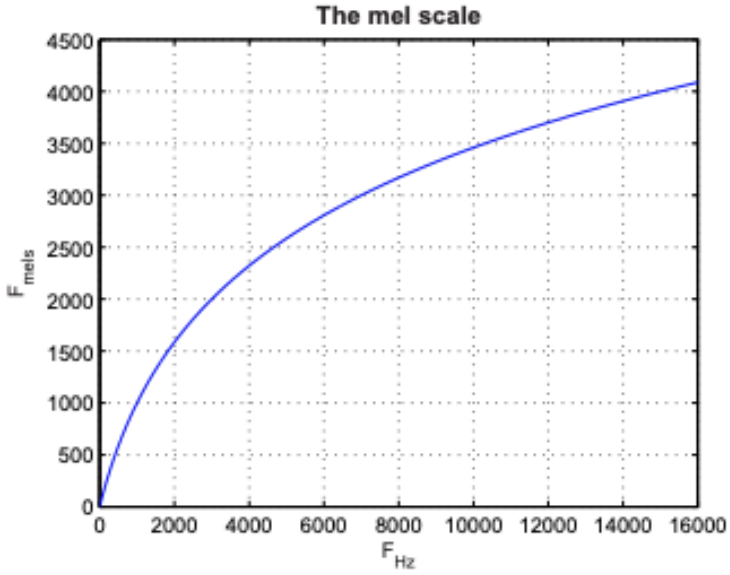

Imagen 13 - Visualización escala Mel [IMG 13]

De acuerdo a los experimentos que se realizaron en los estudios, el tono es linealmente percibido en frecuencias cuyo rango se encuentra entre 0-1000 Hz. Por sobre los 1000 Hz, la escala se vuelve logarítmica.

Los valores de magnitud sobre las bandas de Mel se introducen luego en una Transformada de Coseno Discreta *(Discrete Cosine Transform, DCT),* que produce un espectro calculado sobre las frecuencias de Mel, en lugar de a lo largo del tiempo.

2.3.4. Coeficientes Cepstral de Frecuencia Mel *(Mel Frequency Cepstral Coefficients)*

Los descriptores MFCC corresponden a una simple transformación coseno de la energía, que deriva de la señal de un espectrograma medida en la escala Mel por diferentes sub-bandas filtradas [8].

La Imagen 14 muestra las etapas por las cuales debe pasar una señal de audio para poder extraer los descriptores de audio MFCC.

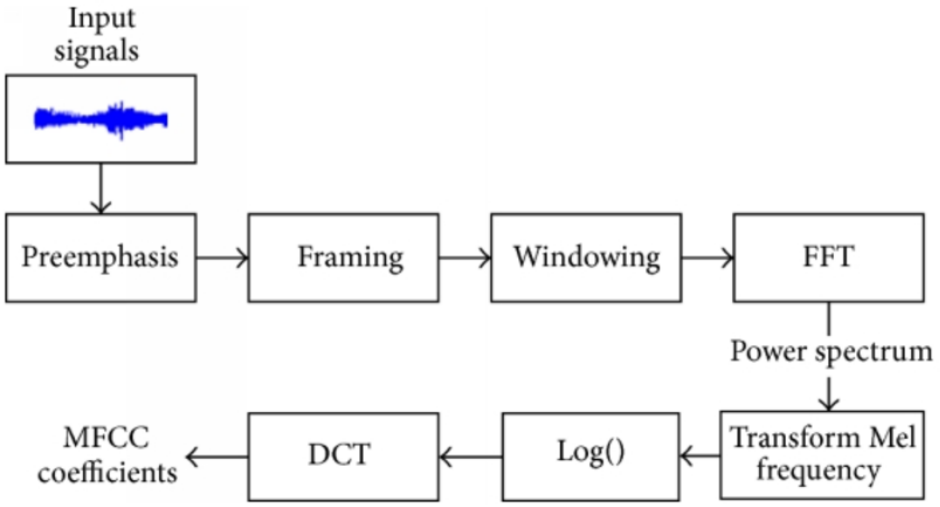

Imagen 14 - Proceso extracción descriptores MFCC [IMG 14]

Para caracterizaciones de audio, los descriptores MFCC entregan buenos resultados para determinar fenómenos relacionados al habla y son muy útiles para reconocer la voz. En la actualidad también son utilizados en soluciones relacionadas a la música, sin perder foco en su origen que está relacionado al procesamiento del habla [4].

Un vector MFCC describe periodicidades que encuentra en los valores de magnitud. Estos valores los extrae desde la distribución de frecuencia de una ventana. Por ejemplo, en el

procesamiento de señales de música, se calculan entre 13 y 25 MFCC normalmente para cada ventana de audio.

El primer MFCC corresponde a la energía promedio de la señal en el cuadro en consideración. Sin embargo, esto no ocurre en el modelado de funciones para la recuperación de música. Las posiciones MFCC crecientes corresponden a periodicidades más altas.

En la Imagen 15, se puede visualizar la forma de onda, el espectrograma y los 20 MFCC correspondientes a lo largo del tiempo.

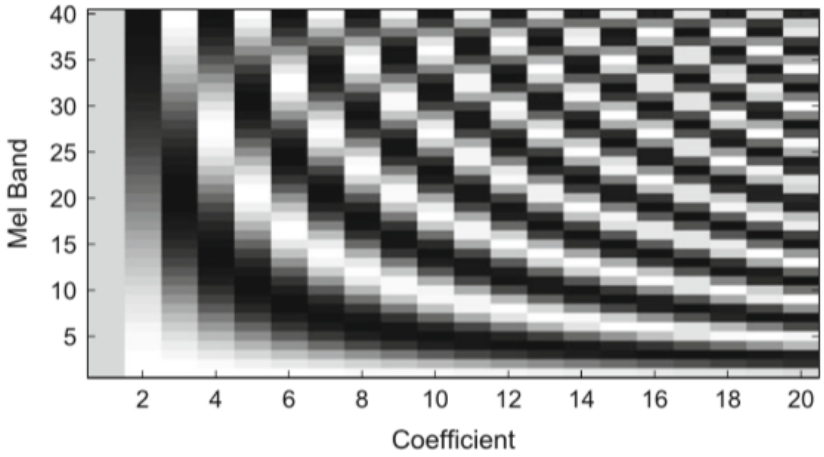

Imagen 15 - Visualización descriptor MFCC [IMG 15]

La computación de vectores MFCC para una pieza musical completa, generalmente produce varias decenas de miles de vectores de características individuales.

Por ejemplo, una canción de 3 minutos muestreada a 44100 Hz, un tamaño de ventana de 512 muestras y un tamaño de salto del 50% produce más de 20000 vectores de características para describirlo, donde el factor de 1.5 explica el tamaño del salto, lo que aumenta el número de cuadros en un 50%.

$$Nfv = 1.5 * \left(\frac{44100\ samples/s}{512\ samples/frame}\right) * 180\ s = 23256 \text{ ventanas}$$

La agregación de este gran número de vectores de características de audio se realiza normalmente mediante el resumen estadístico de todos los vectores resultantes, la aplicación de la cuantización de vectores, o el ajuste de modelos probabilísticos a los datos.

Todos estos métodos descartan el ordenamiento temporal de los *frames*, por lo que proporcionan representaciones de la distribución de los vectores MFCC en toda la pieza.

A partir de la representación resultante de cada canción, las similitudes de pares entre canciones se pueden calcular de varias maneras, según la técnica de agregación.

Si se ejecutaran los pasos detallados anteriormente para todas las ventanas de la pista de audio, se pueden obtener vectores ordenados temporalmente de MFCC, similares al espectrograma producido por una Transformada de Fourier a Corto Plazo *(Short-time Fourier Transform, STFT)*.

2.4. Extracción de Características *(Feature Extraction)*

El propósito de este proceso es obtener, transformar y combinar características originales del contenido multimedia. Existe una gran variedad de técnicas para extraer características, las cuales dependen del propósito que se quiera obtener, como también, del tipo de contenido multimedia desde el cual se extraerán los metadatos.

Los metadatos que se extraen de un objeto multimedia contienen anotaciones, descripciones y características propias del contenido multimedia. Con estos datos se pueden realizar búsquedas de diferente tipo. A continuación, se listan diferentes tipos de metadatos:

- Descriptivos, corresponden a datos relacionados al autor, fecha de creación, largo del objeto, etc.
- Anotaciones, corresponden a descripciones textuales relacionadas al contenido del objeto
- Características, corresponden a características propias del contenido multimedia. Para describir el tipo de características, usualmente se utiliza un tipo de lenguaje llamado MPEG-7 [R14], que corresponde en la actualidad a uno de los más importantes estándares.

El proceso para capturar características de un objeto multimedia se llama Extracción de Características. Este proceso a menudo se realiza automáticamente. Sin embargo, en algunas ocasiones necesita del apoyo humano. Existen dos clases de características, llamadas de bajo nivel y de alto nivel.

2.4.1. Bajo Nivel *(Low Level)*

Las características de Bajo Nivel captan patrones de datos y estadísticas de un objeto multimedia y dependen en gran medida del medio. La extracción de características de Bajo Nivel se realiza automáticamente.

Por ejemplo, para los documentos de texto ¿Qué "contenido" se puede derivar automáticamente? Durante el proceso de indexación, las palabras como "the", "it", "a", etc. se descuidan. Es decir, no tienen ninguna relevancia para el significado del documento. Por lo tanto, el resultado es esencialmente una lista de palabras clave con indicadores de frecuencia, que se supone describe el contenido del documento de texto.

Una señal de audio se puede representar mediante una secuencia de amplitud-tiempo: para cada valor de muestra se cuantifica la presión de aire. La amplitud que pertenece a la presión de aire del silencio se representa como “0”, una presión de aire más alta que la presión de silencio significa una amplitud positiva, y una presión de aire más baja implica una amplitud negativa.

La Transformada Discreta de Fourier (DFT) de la representación de amplitud-tiempo también se utiliza para derivar características de bajo nivel.

Al procesar imágenes, se puede contar el número de píxeles que tienen un color en un rango de color determinado, dando lugar a los llamados histogramas de color. Se pueden usar histogramas de color para distinguir imágenes.

Si una imagen tiene muchos puntos oscuros adyacentes a puntos claros, entonces tiene una puntuación alta con respecto al contraste de la característica. Se han definido muchas otras características de bajo nivel, como por ejemplo: forma, circularidad, dimensionalidad, entre otros.

Los videos son secuencias de imágenes, por lo que las características de bajo nivel de las imágenes también se aplican al video. El video es un medio continuo y tiene como tal una dimensión temporal.

Sea “una toma” como una secuencia de imágenes tomadas con la misma posición de cámara. El final de “una toma” se puede determinar calculando la diferencia de píxeles entre las imágenes posteriores. Tan pronto la diferencia de píxeles entre dos imágenes sea mayor que un cierto umbral, se puede suponer que ha ocurrido un cambio en la toma.

Las características de bajo nivel a menudo no tienen demasiado significado para el usuario final.

2.4.2. Alto Nivel *(High Level)*

Para las imágenes. ¿Qué significa realmente un histograma de color? Mucho verde puede indicar muchas cosas: ¿un campo de golf o tal vez un bosque?

Las características de Alto Nivel (o conceptos de alto nivel) se refieren a características que son significativas para el usuario final, como el bosque o el campo de golf. Hay una brecha entre las características de bajo y alto nivel. Esta brecha se llama brecha semántica. La extracción de características de alto nivel intenta cerrar esta brecha e intenta reconocer conceptos que son significativos para el usuario.

Las características de bajo nivel en un documento de texto son palabras clave. Las palabras clave tienen una fuerte relación con los conceptos en la mente humana. Cuando un documento contiene palabras como "fútbol" y "árbitro", esto proporciona una indicación del contenido del documento.

Para el reconocimiento de voz, en muchos idiomas se han construido traductores razonables de voz a texto. Por ejemplo, para una cierta fuente de datos que contiene voz, los extractores pueden derivar automáticamente características de bajo nivel del habla y, aparentemente, los traductores cierran con éxito la brecha entre las características de bajo nivel por un lado y las palabras y oraciones por el otro.

En áreas como imágenes, audio sin voz y conceptos derivados de video a partir de características de bajo nivel, en general no es posible. Centrarse en un dominio de aplicación especial fomenta el progreso. Por ejemplo, videos de partidos de fútbol. Observar un sonido fuerte proveniente de la multitud y un objeto redondo que pasa una línea blanca, seguido de un silbido agudo, a menudo indica un concepto semánticamente interesante: un gol. Por lo tanto, una combinación de características de bajo nivel puede implicar una característica de alto nivel.

2.4.3. Maldición de la Dimensionalidad *(Curse of Dimensionality)*

En muchas oportunidades el proceso de extracción de características produce vectores que poseen una gran cantidad de dimensiones, incrementando la complejidad en su manejo y procesamiento de resultados.

La Maldición de la Dimensionalidad se refiere al fenómeno que puede perjudicar la calidad de los resultados que se quieren obtener, producto de la dispersión de los datos en el espacio y la complejidad de controlar el ruido en los datos, resultando en una ejecución lenta de procesar y con baja precisión en los resultados entregados [4].

Una de las principales preocupaciones corresponde al número de posibles configuraciones que se pueden obtener en un conjunto de variables, la cual crece exponencialmente a medida que aumenta el número de dimensiones. Es decir, a medida que aumenta el número de dimensiones de datos (de izquierda a derecha, ver Imagen 16), el número de configuraciones de interés podría crecer exponencialmente.

Al tener solamente una dimensión, solo importa distinguir las 10 celdas que posee la variable. Con la cantidad de ejemplos suficientes para cubrir cada una de las celdas, los algoritmos pueden fácilmente diferenciar los datos correctamente.

Con 2 dimensiones, resulta más complejo diferenciar 10 valores diferentes para cada una de las variables. En este caso, correspondería a 10x10 = 100 celdas que necesitan la cantidad de ejemplos suficientes para ser cubiertas con datos. Si fueran 3 dimensiones, correspondería a 10x10x10 = 1000 celdas.

Por lo tanto, a medida que aumenta el número de dimensiones, se requieren más datos para cubrir la cantidad de celdas de cada variable.

Imagen 16 - Maldición de la dimensionalidad [IMG 16]

2.5. Reducción de Dimensiones *(Dimensionality Reduction)*

Reducir dimensiones tiene como propósito identificar las características que son más relevantes para realizar una tarea, permitiendo obtener una simplificación más representativa de los patrones que se buscan identificar, disminuyendo la memoria utilizada y el tiempo de ejecución.

Al tener grandes cantidades de dimensiones de datos, buscar e identificar similitudes se vuelve más complicado producto del procesamiento que se debe realizar. Por lo tanto, este

método tiene como objetivo reducir el número de características obtenidas en la representación vectorial de un objeto multimedia.

2.6. Videos

Si bien las imágenes tienen un alto y ancho, un modelo de colores y valores en cada píxel, los videos corresponden a un conjunto de imágenes secuenciales, las cuales al ser reproducidas usualmente generan una acción de movimiento (ver Imagen 17).

Existen propiedades adicionales en los videos que son necesarias para obtener los descriptores de imagen, como los FPS *(frames per second)*.

Los denominados FPS corresponden a la cantidad de cuadros (imágenes) que se deben mostrar en cada segundo. Mientras más FPS tenga un video, más fluido será el movimiento del contenido que reproduce.

Es importante considerar los FPS de un video, porque están directamente relacionados con la cantidad de datos que se pueden llegar a extraer para crear los descriptores de imagen. Es decir, extraer ~30 imágenes por segundo puede que no sea una buena estrategia en términos de rendimiento y tamaño del descriptor.

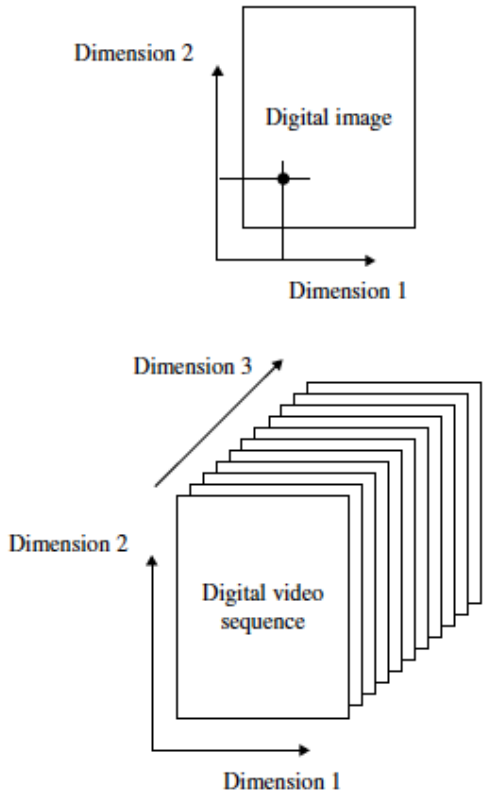

Imagen 17 - Representación secuencia de imágenes en video [IMG 17]

Para obtener un descriptor robusto y capturar la mayor cantidad de datos utilizando pocas imágenes, basta extraer una cantidad reducida de cuadros por segundo, disminuir el tamaño de la imagen y realizar operaciones matemáticas para obtener un descriptor único.

La Imagen 18 muestra 3 imágenes obtenidas en diferentes momentos de la secuencia del video. En ella, se puede visualizar que tan solo algunos valores de los píxeles cambia, mientras que el resto de los valores se mantiene.

Imagen 18 - Extracción de características visuales

2.7. Espectrograma *(Espectrogram)*

Corresponde a una representación visual del espectro de frecuencias de una señal que varía en el tiempo. Su objetivo busca entender cómo la energía de las frecuencias varía en el tiempo.

El cambio de energía de una frecuencia en el tiempo es representado por una tercera dimensión de colores. Los azules oscuros corresponden a bajas amplitudes, mientras que colores más claros, como el rojo, corresponden a señales con mayor energía.

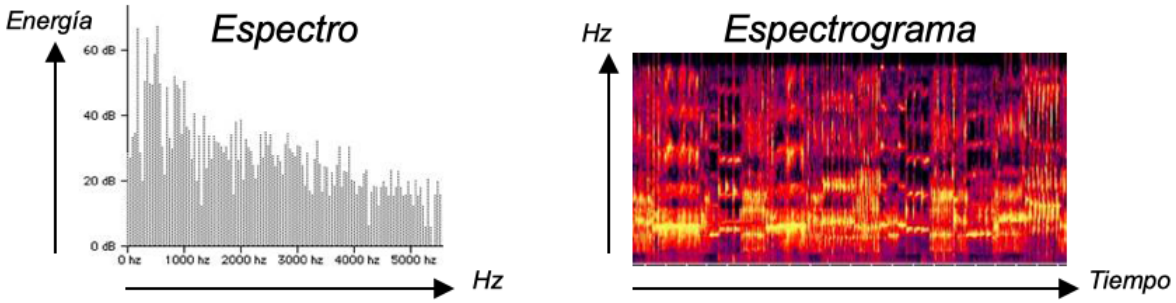

Imagen 19 - Visualización espectrograma

2.8. Búsqueda de Vecinos Cercanos *(Nearest Neighbors Search)*

Corresponde a uno de los algoritmos más simples de utilizar y es conocido por realizar una búsqueda exacta de vecinos cercanos.

Su funcionalidad se basa en identificar un número finito de objetos cercanos *"k"* a una consulta *Q,* medidos en base a una función de distancia establecida que generalmente es Euclidiana.

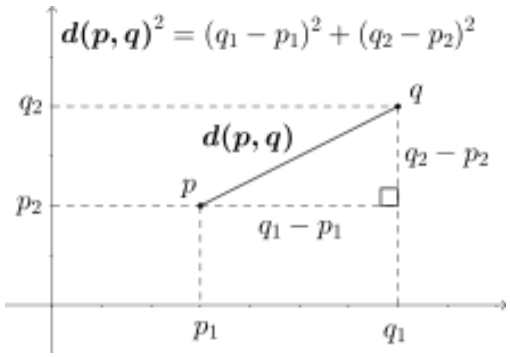

Imagen 20 - Distancia Euclidiana

Este algoritmo comúnmente puede ser utilizado en métodos de aprendizaje supervisado, y a diferencia de otros modelos, es conocido por ser *lazy learner*, puesto que en vez de aprender los datos que tiene de entrenamiento, tan solo los recuerda [1].

La Imagen 21 permite visualizar cómo se identifican los cinco vecinos más cercanos, respecto a una consulta posicionada al centro de la circunferencia.


Imagen 21 - Búsqueda de vecinos cercanos

Cuando los descriptores tienen una gran cantidad de dimensiones, como ocurre con el contenido multimedia, la Búsqueda de Vecinos Cercanos puede que no sea la mejor

alternativa, producto del tiempo que demora el algoritmo en realizar el cálculo de distancias entre los vectores y la obtención de resultados.

2.8.1. Búsqueda Aproximada

La idea general de los algoritmos de búsqueda aproximada es disminuir algunas restricciones en la búsqueda de similitud exacta, para reducir los costos de búsqueda medidos por los accesos a disco y/o el número de cálculos de distancia.

Esto significa que se pueden producir resultados que no son correctos, o bien, que no son identificados en la búsqueda.

Como se ha expuesto en la Búsqueda de Vecinos Cercanos, este tipo de búsqueda es muy utilizada en la actualidad y se justifica principalmente para los siguientes escenarios:

(1) La similitud entre los objetos es a menudo subjetiva, por lo que es muy difícil de expresar mediante una función rigurosa y única. Para una base de datos de imágenes, dada una imagen de consulta y un conjunto de resultados candidatos de imágenes, diferentes personas elegirían distintas imágenes de acuerdo a la consulta realizada.

Cuando la noción intuitiva de similitud se define formalmente por una fórmula matemática (función de distancia), la subjetividad no se tiene en cuenta. Por lo tanto, la imprecisión en el resultado de búsqueda podría ser tolerada por los usuarios.

(2) Los procesos de búsqueda de similitud son intrínsecamente iterativos. Los usuarios suelen emitir varias consultas en el sistema de búsqueda, posiblemente reutilizando resultados de consultas anteriores para ejecutar otras nuevas. Un usuario puede realizar una búsqueda utilizando una imagen inicial como consulta para encontrar imágenes similares. Al no estar satisfecho con el resultado, el usuario puede ejecutar otra consulta de búsqueda de similitud utilizando una de las imágenes devueltas previamente como referencia. Con este enfoque, es importante una ejecución eficiente de consultas elementales y los usuarios pueden aceptar cierta imprecisión en los resultados temporales, siempre que la ejecución de consultas sea rápida

Para la mayoría de los casos, se ha demostrado que en las aplicaciones más comunes las búsquedas aproximadas realizan una buena tarea de aproximación para listar resultados, ordenando los datos mucho más rápido que librerías estándar al ejecutar la misma consulta.

Una alternativa para ejecutar búsquedas aproximadas corresponde al algoritmo FLANN *(Fast Library for Approximate Nearest Neighbors)* que realiza el cálculo de distancias aproximadas rápidamente en espacios compuestos por muchas dimensiones, e identifica los vectores más cercanos de acuerdo a un valor "*k*" parametrizable.

Al igual que la búsqueda exacta de vecinos cercanos, FLANN puede calcular la distancia entre vectores utilizando, por ejemplo, la distancia Euclidiana.

Para ordenar los datos rápidamente, como estructura de datos utiliza árboles binarios llamados *kd-tree*, los cuales indexan los descriptores visuales y de audio que se generen (ver Imagen 22).

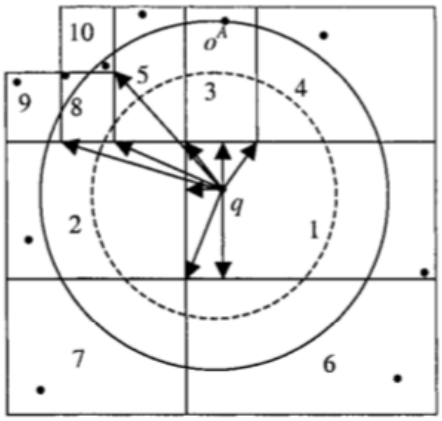

Imagen 22 - Búsqueda aproximada de vecinos cercanos [IMG 22]

Los *kd-tree* corresponden a una estructura de datos utilizada para organizar cierto número de puntos en un espacio con "*k*" dimensiones. Este tipo de árboles resulta muy útil para las búsquedas por rango o de vecinos más cercanos.

Cada nivel de un árbol *k-d* divide todos los elementos secundarios a lo largo de una dimensión específica, utilizando un hiperplano que es perpendicular al eje correspondiente.

En la raíz del árbol, todos los elementos secundarios son divididos según la primera dimensión. Es decir, si la primera coordenada de dimensión es menor que la raíz, estará en el subárbol de la izquierda, pero si es mayor que la raíz, estará en el subárbol derecho.

Para decidir cuál dimensión será dividida, primero se deben calcular las varianzas para cada una de las dimensiones. Luego, se elige la dimensión con mayor varianza y se divide

el conjunto en dos, según el valor de la mediana, y continúa dividiendo recursivamente los sub conjuntos que aparezcan.

Existen dos tipos de nodos: Los internos, que se encargan de almacenar un número de dimensión y el valor del umbral. Y por otro lado los externos, que se encargan de almacenar los vectores.

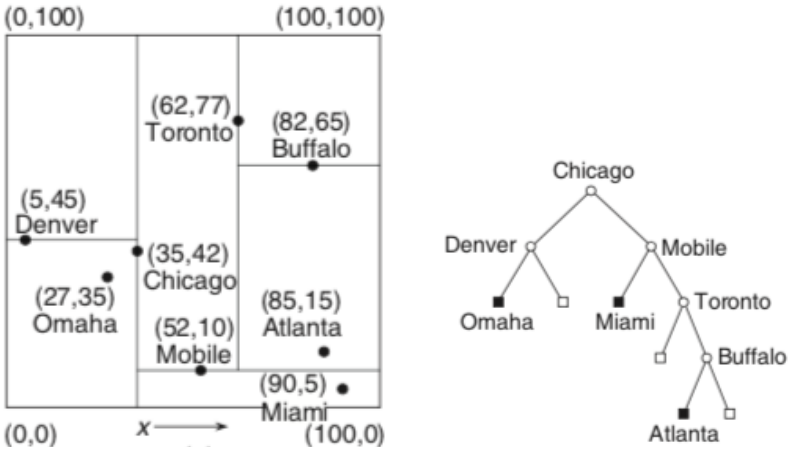

Imagen 23 - Representación del árbol k-d [IMG 23]

2.9. Evaluación de Conjuntos de Recuperación

Para medir la efectividad de un sistema, existen dos indicadores dentro del área de Recuperación de Información, denominadas *Precision* y *Recall*.

La Imagen 24 muestra cómo a partir de una colección de documentos se pueden identificar aquellos que son relevantes, dada una intersección $|Ra|$ entre el set de documentos relevantes $|R|$ y el set de respuestas $|A|$.

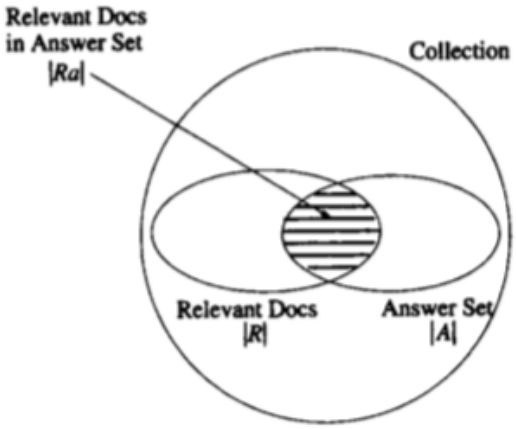

Imagen 24 - Precision y Recall [IMG 24]

Precision (P) corresponde a la fracción de documentos recuperados que son relevantes. Es una relación del número de verdaderos positivos, dividida por la suma de los verdaderos positivos y los falsos positivos.

Su objetivo busca responder a la siguiente pregunta:
¿Cuál es la proporción de identificaciones correctas?

$$P = \frac{relevant\ items\ retrieved}{retrieved\ items} = \frac{relevant}{retrieved}$$

$$P = \frac{|Ra|}{|R|}$$

Recall (R) corresponde a la fracción de documentos relevantes que son recuperados. Se calcula como la proporción del número de verdaderos positivos, dividida por la suma de los verdaderos positivos y los falsos negativos.

Su objetivo busca responder a la siguiente pregunta:
¿Cuál es la proporción total identificada de positivos reales?

$$R = \frac{relevant\ items\ retrieved}{relevant\ items} = \frac{retrieved}{relevant}$$

$$R = \frac{|Ra|}{|A|}$$

Precision y *Recall* asumen que todos los documentos del set de datos son examinados. Sin embargo, al usuario no se le presentan todos los resultados, dado que son ordenados de acuerdo a un grado de relevancia. A medida que el usuario continúa con la examinación del set de respuestas, las medidas de *Precision* y *Recall* varían.

Muchas veces un sistema de Recuperación de Información es juzgado por su *Precision*. Es decir, la fracción de sus recuperaciones que son correctas.

Estas nociones se pueden aclarar al examinar la siguiente matriz de confusión:

	Relevantes	**No Relevantes**
Recuperados	Verdaderos Positivos (TP)	Falsos Positivos (FP)
No Recuperados	Falsos Negativos (FN)	Verdaderos Negativos (TN)

Tabla 1 - Matriz de Confusión

Por lo tanto,

$$P = TP/(TP + FP)$$
$$R = TP/(TP + FN)$$

En términos de la matriz de confusión anterior, la tasa de correctas *Accuracy (A)* se podría calcular como:

$$A = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

Los sistemas de Recuperación de Información pueden considerarse como un clasificador de dos clases que intenta etiquetar los documentos como relevantes y no relevantes. Esta es precisamente la medida de efectividad que se usa a menudo para evaluar problemas de clasificación de aprendizaje automático.

Una buena razón por la cual *Accuracy* no es una medida adecuada para evaluar sistemas de Recuperación de Información, corresponde a que en casi todas las circunstancias los

datos son extremadamente sesgados. Normalmente más del 99,9% de los documentos se encuentran en la categoría no relevante.

Un sistema ajustado para maximizar *Accuracy* puede parecer que funciona bien simplemente considerando que todos los documentos no son relevantes para todas las consultas. Incluso si el sistema es bastante bueno, tratar de etiquetar algunos documentos como relevantes casi siempre conducirá a una alta tasa de falsos positivos. Por otro lado, etiquetar todos los documentos como no relevantes es completamente insatisfactorio para un usuario del sistema de Recuperación de Información.

Los usuarios siempre querrán ver algunos documentos. Sin embargo, se puede suponer que tienen cierta tolerancia para ver algunos falsos positivos, siempre que obtengan información útil.

Las medidas de *Precision* y *Recall* concentran la evaluación en el retorno de verdaderos positivos, preguntando qué porcentaje de los documentos relevantes se han encontrado y cuántos falsos positivos también se han devuelto.

La ventaja de tener valores para *Precision* y *Recall,* es que uno es más importante que el otro en muchas circunstancias. Por ejemplo, los usuarios habituales de algún buscador en la web desean que todos los resultados de la primera página sean relevantes (*Precision*), pero no tienen el menor interés en saber y mucho menos en mirar las siguientes páginas.

Existen otras aplicaciones que requieren obtener *Recall* tan alto como sea posible, y tolerarán resultados de *Precision* bastante bajos para obtener documentos. Por ejemplo, sistemas que detectan enfermedades, sismos, violaciones de seguridad o cualquier otro contexto de características que involucren una urgencia, es preferible que entreguen un resultado.

Recall es una función no decreciente del número de documentos recuperados. En un buen sistema, *Precision* generalmente disminuye a medida que aumenta el número de documentos recuperados.

Con frecuencia hay *trade-off* entre los datos que generan los valores de *Precision* y *Recall*. En definitiva, lo que usualmente se busca es obtener una cierta cantidad de datos recuperados, mientras se tolera solo un cierto porcentaje de falsos positivos.

Teniendo en cuenta que *Precision* y *Recall* son medidas basadas en conjuntos, se calculan utilizando documentos no ordenados.

Para evaluar completamente la efectividad de un modelo, se debe examinar cual es el comportamiento de los valores de *Precision* y *Recall*.

En un contexto de recuperación, los conjuntos apropiados de documentos recuperados están naturalmente dados por la parte *"k"* superior de los documentos obtenidos. Para cada conjunto de este tipo, los valores de *Precision* y *Recall* se pueden trazar para generar una curva de *Precision* y *Recall*.

Una curva de *Precision* y *Recall*, es una gráfica de *Recall* (eje x) y *Precision* (eje y). Usualmente se aplica en base a 11 niveles estandarizados de *Recall* que corresponden a los valores 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%.

Este tipo de curva tiene una forma distintiva de diente de sierra: si el valor *(k + 1)* del documento recuperado no es relevante, la recuperación es la misma que para los documentos *"k"* superiores a él, disminuyendo *Precision*. Por el contrario, si fuese relevante, tanto *Precision* y *Recall* aumentan, desplazando la curva hacia la derecha [R11].

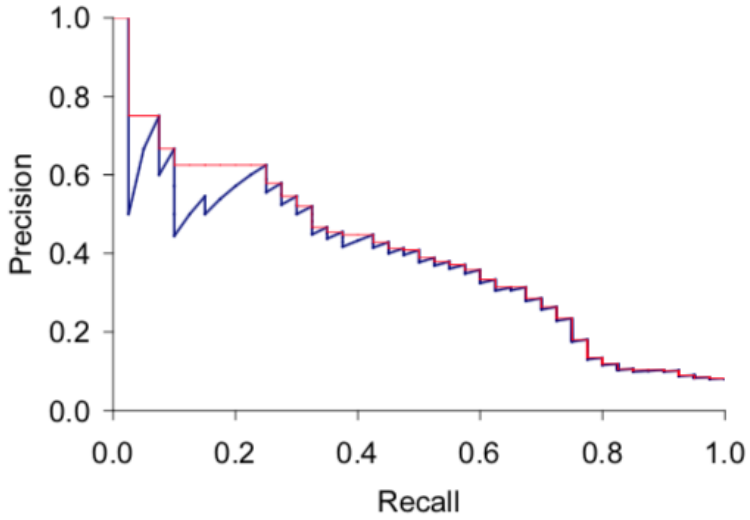

Imagen 25 - Curvas de Precision y Recall [IMG 25]

2.9.1. Mean Average Precision *(MAP)*

Corresponde a una medida que combina *Precision* y *Recall* para obtener resultados de recuperación de documentos relevantes.

Esta medida se calcula como el promedio de *Precision*, denominado *Average Precision (AP)*, en el rango de cada documento relevante recuperado. A los documentos relevantes que no se recuperan se les asigna un valor *Precision* de cero.

Por ejemplo, para una consulta *Q* que busca obtener como resultado la palabra *"auto"*, el sistema de Recuperación de Información entrega los siguientes resultados:

auto, bicicleta, auto, auto, bus, camioneta, auto

El cálculo *AP* corresponde al siguiente:

$$AP = \left(\frac{1}{1} + 0 + \frac{2}{3} + \frac{3}{4} + 0 + 0 + \frac{4}{7}\right)/4 = 0.747$$

Si se realizara este ejercicio para diferentes consultas, se obtendrían diferentes *AP* para cada una de ellas.

La media de todos los puntajes *AP* calculados entregan un solo resultado llamado *MAP*, que cuantifica qué tan bueno es el modelo para realizar una consulta y entregar resultados relevantes.

$$MAP = \sum_{q=1}^{Q} \frac{AveP(q)}{Q}$$

3. Metodología de Trabajo

Dada la necesidad expuesta como problema y el desconocimiento de las variaciones que puedan surgir en la construcción de esta solución, en el proceso de desarrollo se utilizará *XP (extreme programming)*, bajo un ciclo de vida iterativo incremental. Esta selección híbrida permitirá sostener y dar solución a las variaciones que pueda necesitar la herramienta frente a las observaciones o cambios que pueda realizar el cliente.

Como se utilizarán prácticas y tecnologías en las cuales se requerirá invertir tiempo para su investigación, hacer uso de esta selección se considera adecuada para finalizar el proyecto exitosamente.

Se evaluará la calidad de los resultados obtenidos, de acuerdo a la cantidad de aciertos o errores que la detección genere.

La metodología seleccionada estará compuesta por las siguientes actividades:

1. Documentación tesis
    - Iteración I
        - Planificación
        - Investigación
        - Implementación de un módulo para descargar televisión online
            - Codificación
            - Pruebas
        - Implementación de un algoritmo de detección basado en características visuales
            - Codificación
            - Pruebas
    - Iteración II
        - Planificación
        - Investigación
        - Implementación de un algoritmo de detección basado en características de audio
            - Codificación
            - Pruebas
    - Iteración III
        - Planificación
        - Investigación
        - Implementación de un algoritmo de detección basado en características audiovisuales
            - Codificación
            - Pruebas
    - Evaluación de resultados
2. Conclusiones
3. Escribir informe de tesis

Imagen 26 - Iteraciones sobre la metodología

4. Plan de Trabajo

De acuerdo a la metodología seleccionada, la ejecución del proyecto estará guiada por una planificación de trabajo a realizar en ocho meses continuos de duración, dentro de los cuales, la primera mitad del tiempo estará dirigido al prototipado de las iteraciones I y II.

Por otra parte, la segunda mitad del trabajo se concentrará en la iteración III que finalizará concluyendo, de acuerdo a los resultados obtenidos, cuál de todos los prototipos entregó mayor precisión y recuperación en sus resultados.

	Iteración I		**Iteración II**		**Iteración III**			
	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Mes 7	Mes 8
Planificación	**X**		**X**		**X**			
Investigación y Codificación	**X**	**X**	**X**	**X**	**X**	**X**	**X**	
Prototipado y Pruebas		**X**		**X**	**X**	**X**	**X**	**X**

Evaluación de Resultados		X		X				X
Documentación		X		X	X	X	X	X

Tabla 2 - Plan de trabajo

5. Definición y Alcance

Dado los objetivos planteados en el alcance de este proyecto de tesis, se desarrollarán scripts para ser ejecutados en un entorno *bash* que permitirán realizar la descarga de las grabaciones de comerciales desde los servicios de streaming que proveen los canales locales de televisión.

Por otro lado, la generación de estadísticas se podrá ejecutar mediante scripts montados en *Jupyter*, los cuales permitirán soportar las visualizaciones requeridas en las tareas descritas para reportería y *dashboards*.

Las tareas de detección, para las diferentes modalidades desarrolladas, se podrán ejecutar en un entorno web montado en un equipo con al menos 2.4 GHz de procesamiento y 8 Gb en memoria RAM.

Imagen 27 - Objetivos dentro del alcance

6. Etapa de Solución Técnica

De acuerdo con el alcance de este proyecto, la solución técnica solo estará enfocada en el desarrollo *back-end* de la solución, basado en scripts encargados de descargar el contenido multimedia desde los servicios de streaming mencionados anteriormente. Cada archivo multimedia será procesado para extraer los descriptores de imagen y de audio necesarios para realizar una tarea de detección.

Teniendo en cuenta que ejecutar tareas de Recuperación de Información y detección de similitudes son de alta complejidad técnica, se investigarán diferentes alternativas, mencionadas en el marco teórico, con el propósito de seleccionar la que se adapte de mejor forma para cubrir los objetivos y responder las hipótesis planteadas para este proyecto.

Como el objetivo de este proyecto es identificar las similitudes más cercanas de acuerdo al cálculo de distancias, no se entrenarán modelos de aprendizaje para cubrir la necesidad.

En los resultados de evaluación, las alternativas seleccionadas permitirán cubrir el alcance y ayudarán a finalizar el proyecto en el rango de tiempo estipulado en el plan de trabajo.

6.1. Proceso de Detección

Para que una detección de comerciales sea exitosa, es necesario ejecutar una serie de tareas previas que se pueden listar en el siguiente orden:

1. Obtención de descriptores
2. Cálculo de distancias entre descriptores
3. Identificación de candidatos (vecinos más cercanos)
4. Descartar candidatos

Con respecto a los descriptores, su obtención está compuesta por la creación de vectores de datos calculados desde una imagen, o una pista de audio. Cada vector puede registrar diferentes dimensiones, dependiendo de los datos que se extraen desde el contenido multimedia, como también, del propósito para el cual se requiera utilizar el descriptor.

Dicho lo anterior, extraer características desde un objeto y distribuir sus valores en un vector, es una de las principales tareas para identificar similitudes entre diferentes objetos.

Para calcular la distancia entre vectores e identificar su similitud, se debe utilizar una función de distancia, como la distancia Euclidiana, y definir un umbral cuyo propósito es

descartar aquellos vectores con distancias que se encuentren muy lejanas a un vector de la consulta *Q*.

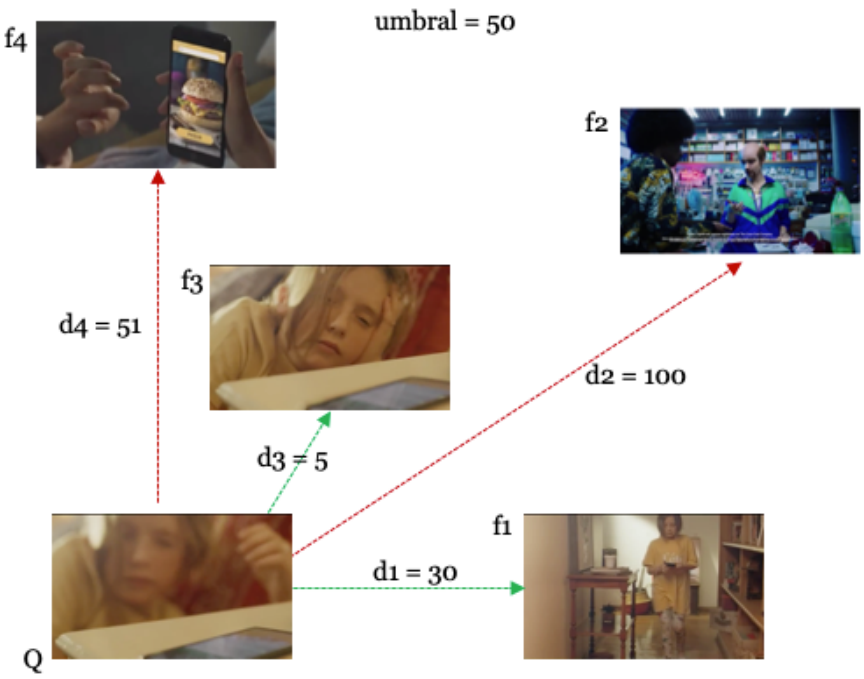

Imagen 28 - Distancia entre Objetos

De acuerdo a la Imagen 28, una consulta Q *(frame)* puede identificar diferentes candidatos luego de haber ejecutado el cálculo de distancia entre vectores. Como las distancias d2 y d4 son superiores al umbral, los vectores f2 y f4 son descartados. Sin embargo, como las distancias d1 y d3 se encuentran dentro del umbral, existen mayores posibilidades para Q, f1 y f3 de compartir valores similares en contenido.

Para ejecutar el cálculo de distancias entre vectores e identificar candidatos rápidamente, se utilizó un algoritmo de búsqueda aproximada como FLANN, el cual calcula la distancia entre vectores, organiza e indexa los resultados en árboles, tales como *kd-tree*.

Es importante considerar que el desarrollo de esta solución solo utiliza un vecino cercano *(k=1)*. Por lo tanto, para cada consulta el algoritmo identifica el vecino cercano con la menor distancia calculada que cumpla con el umbral definido.

Para el ejemplo expuesto anteriormente, y considerando la cantidad de vecinos *k=1*, el vector con menor distancia a *Q* corresponde a *f3*.

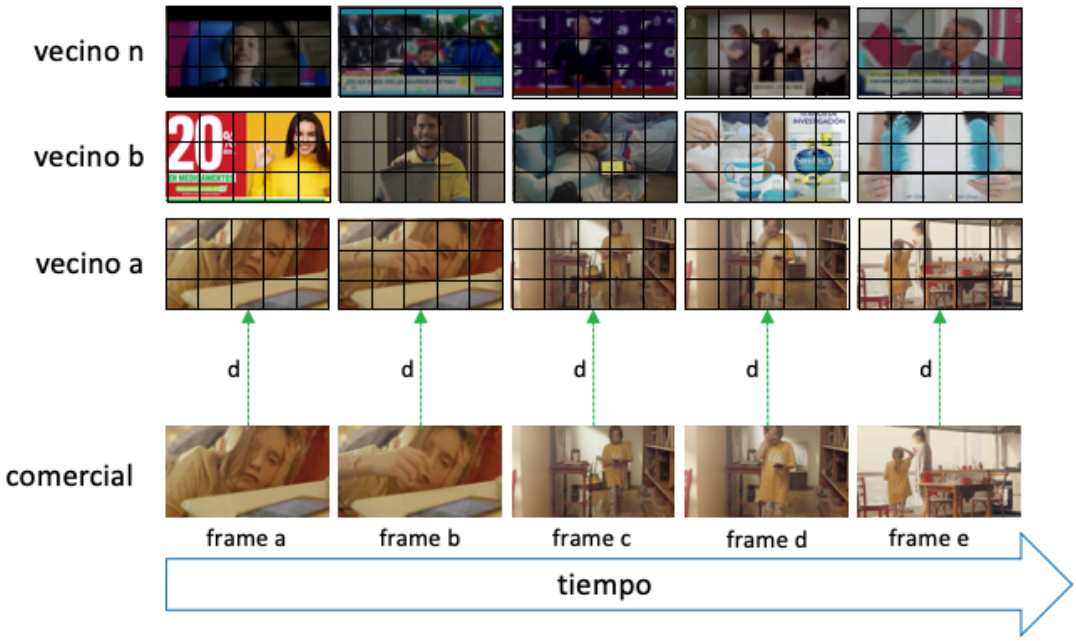

Imagen 29 - Detecciones por cada *frame*

Para que una detección no sea descartada, debe tener a su alrededor *frames* de similares características que también hayan podido identificar a su vecino más cercano dentro del umbral definido. Si la detección no cumple con esta condición, se descarta y el algoritmo continua con el siguiente *frame*.

Una vez identificada cierta cantidad de *frames* consecutivos que cumplan con los criterios mencionados anteriormente, se agrupan los datos obtenidos y la detección se da por completa.

Al completar la detección para un grupo de *frames*, el algoritmo continúa en la búsqueda de similitudes sobre el resto de los datos disponibles, hasta recorrer la base de datos por completo.

6.2. Obtención de Descriptores

Como desde el contenido audiovisual se pueden extraer descriptores de imagen y de audio, las características entre uno y otro son muy diferentes respecto a los datos que almacenan, como también, a la cantidad de dimensiones.

Desde cada video solo se extrajeron tres imágenes por segundo. Cada imagen fue convertida desde su formato a color en un formato a escala de grises.

Una de las principales razones por la cual fue seleccionado utilizar tres imágenes por segundo, corresponde a lo expuesto en la sección de videos 2.6 del marco teórico, donde se muestra que en una fracción de tiempo los valores por píxel de los *frames* no varían mucho. Por esta razón, se creará un único descriptor con el promedio de los valores de los tres descriptores obtenidos por cada segundo.

La Imagen 30 muestra cómo, para una secuencia de video, se obtienen tres *frames* por segundo, de los cuales se extraen los valores para cada píxel y se promedian para generar un único descriptor por segundo.

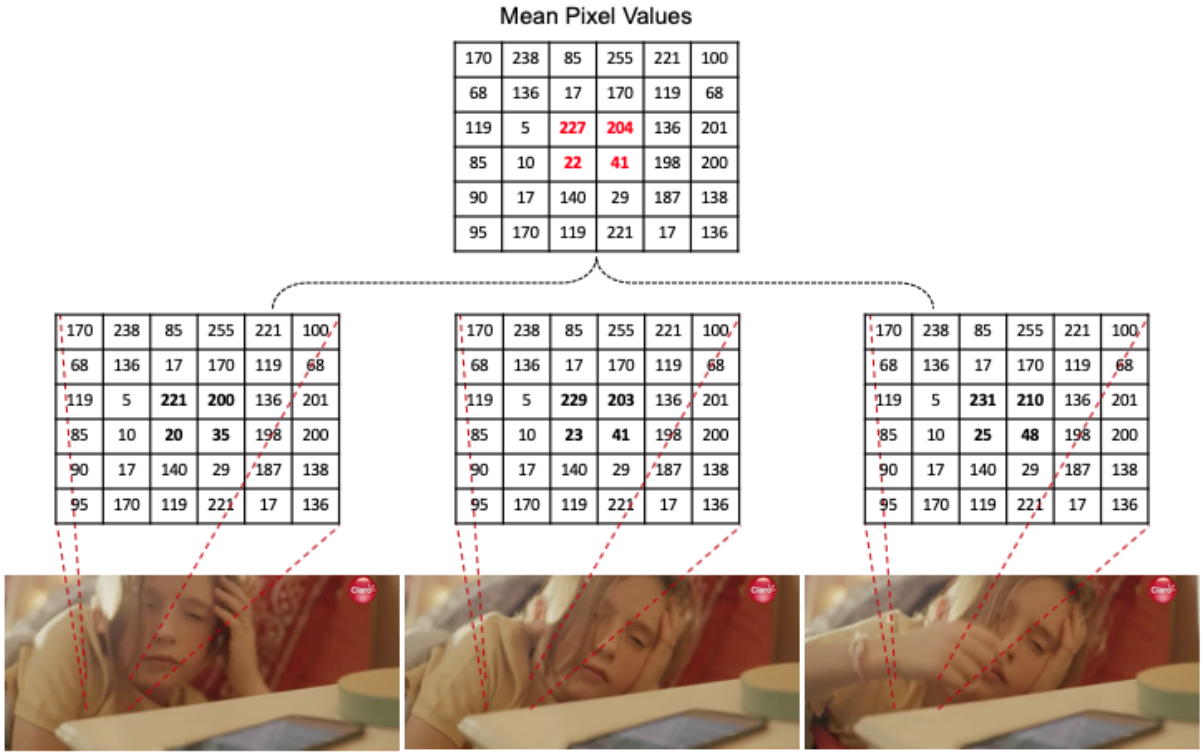

Imagen 30 - Descriptor de imagen promedio de 3 fps

Para responder la segunda hipótesis, se hará uso de la técnica *late fusion*, que consiste en utilizar cada descriptor por separado y luego combinar los resultados de detección obtenidos por cada uno.

A continuación, la siguiente tabla muestra los tipos de descriptores [11, P622] que serán utilizados, junto a las características que los componen:

Descriptor	Tipo	Dimensiones	Tamaño	Duración video
Imagen	10x10 px, mean 3 fps	~(60000, 100)	~50 mb	16 hrs. 40 min
Audio	22050 Hz *sample rate*, 8820 *hop length* + 40 MFCC	~(150000, 40)	~40 mb	16 hrs. 40 min

Tabla 3 - Características de los descriptores

La Imagen 31 muestra cómo se visualizarán las detecciones de cada comercial, expuestos en las siguientes secciones, en las cuales el algoritmo de búsqueda aproximada intenta identificar el *frame* más cercano y de mayor similitud que encuentre.

Imagen 31 - Detección de comercial

Las secciones listadas a continuación, permitirán ejemplificar de mejor manera cómo funciona la búsqueda aproximada utilizando diferentes descriptores para realizar una detección de contenido audiovisual.

A continuación, se presenta la interfaz desarrollada para mostrar cómo funciona la detección de comerciales en un entorno web.

El usuario podrá seleccionar el día que quiere consultar, como también, cargar un archivo multimedia para realizar la tarea de detección pulsando el botón *detect*. El botón *refresh*, tiene la funcionalidad para actualizar la página y liberar memoria.

Los resultados serán visibles mediante la aparición de reproductores multimedia. Es decir, una vez que el algoritmo culmine su ejecución, aparecerán reproductores por cada detección identificada en las diferentes instancias de tiempo en las que se encuentre dentro del archivo multimedia.

6.2.1. Detección Utilizando Descriptor de Imagen

Cuando se selecciona el tipo de detección *image,* los resultados que se obtendrán serán únicamente utilizando el descriptor de imagen.

La Imagen 32 muestra el resultado de realizar una consulta de un comercial para un día en específico, donde se detectaron tres apariciones en distintos momentos del día.

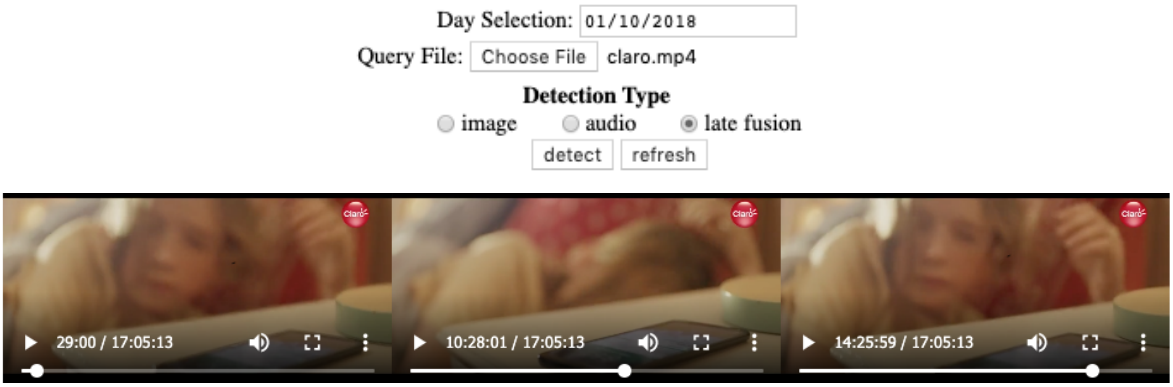

Imagen 32 - Detección utilizando descriptor de imagen

6.2.2. Detección Utilizando Descriptor de Audio

Al seleccionar el tipo de detección *audio,* los resultados que se obtendrán serán únicamente utilizando el descriptor de audio.

En el caso de la detección de audio, la Imagen 33 muestra el resultado de realizar una consulta de un comercial para un día en específico donde los resultados recuperados fueron cuatro.

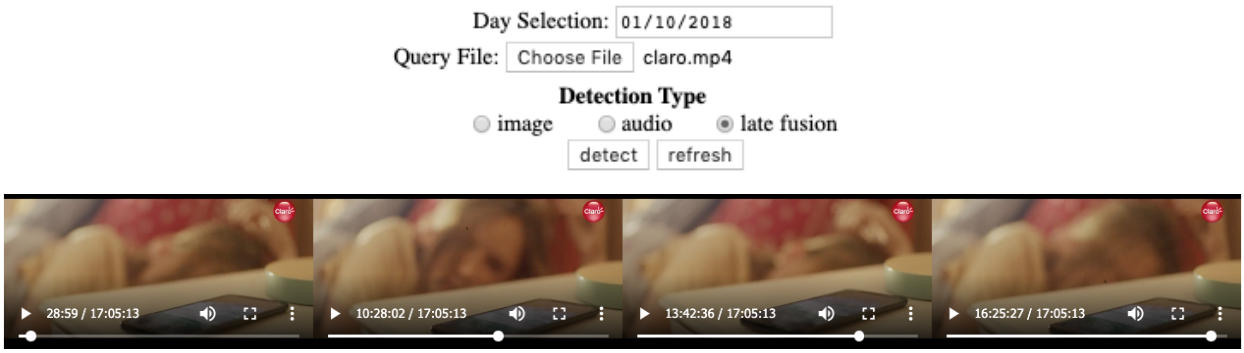

Imagen 33 - Detección utilizando descriptor de audio

6.2.3. Detección Utilizando Descriptor Late Fusion

Al seleccionar el tipo de detección *late fusion,* los resultados que se obtendrán serán utilizando la combinación de descriptores de imagen y audio.

Para la detección *late fusion*, la Imagen 34 muestra el resultado de realizar una consulta de un comercial para un día en específico, donde los resultados recuperados fueron siete.

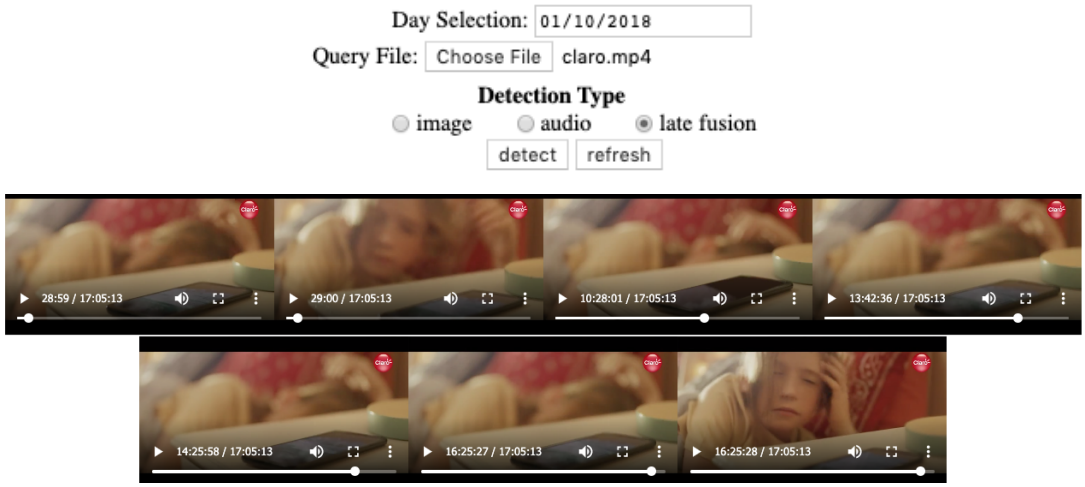

Imagen 34 - Detección utilizando descriptor *late fusion*

7. Etapa de Exploración

Esta sección trata sobre la exploración de los datos que fueron obtenidos por las detecciones ejecutadas en el set de datos de pruebas.

Es importante considerar que, del universo de emisiones identificadas, en algunas visualizaciones solo se hará uso de atributos como las etiquetas, fechas, días de la semana o nombre de los comerciales más representativos y que registren la mayor cantidad de resultados.

7.1. Visualización de Barras

La Imagen 35 muestra un gráfico de barras utilizando el 100% de los datos, los cuales fueron clasificados en 16 etiquetas definidas en la sección de objetivos, con el propósito de identificar tendencias por grupos de productos.

La siguiente tabla muestra las estadísticas obtenidas para este análisis.

count	mean	std	min	25%	50%	75%	max
16	468.25	504.8	42	109.75	269.5	530.25	1688

Tabla 4 - Cantidad de comerciales agrupados por etiqueta

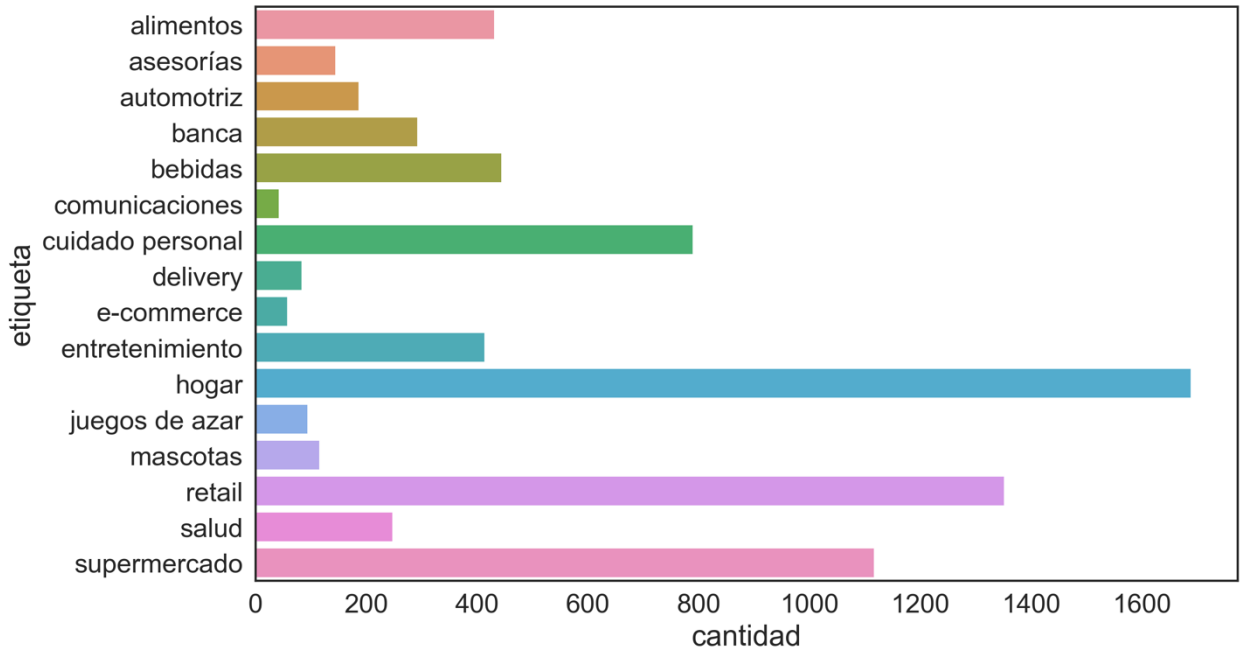

Imagen 35 - Cantidad de comerciales por etiqueta

En resumen, se puede visualizar una alta presencia de comerciales cuya etiqueta corresponde a hogar, donde su contenido está relacionado a productos de decoración, materiales de construcción, limpieza, menaje, aromatización ambiental, entre otros.

7.2. Diagrama de Cajas

A continuación, la Imagen 36 muestra un diagrama de cajas que agrupa el 100% de los datos por fecha y etiqueta.

La siguiente tabla muestra las estadísticas obtenidas para este análisis.

count	mean	std	min	25%	50%	75%	max
402	18.6	21.4	1	4	11	23	127

Tabla 5 - Cantidad de comerciales agrupados por etiqueta y fecha

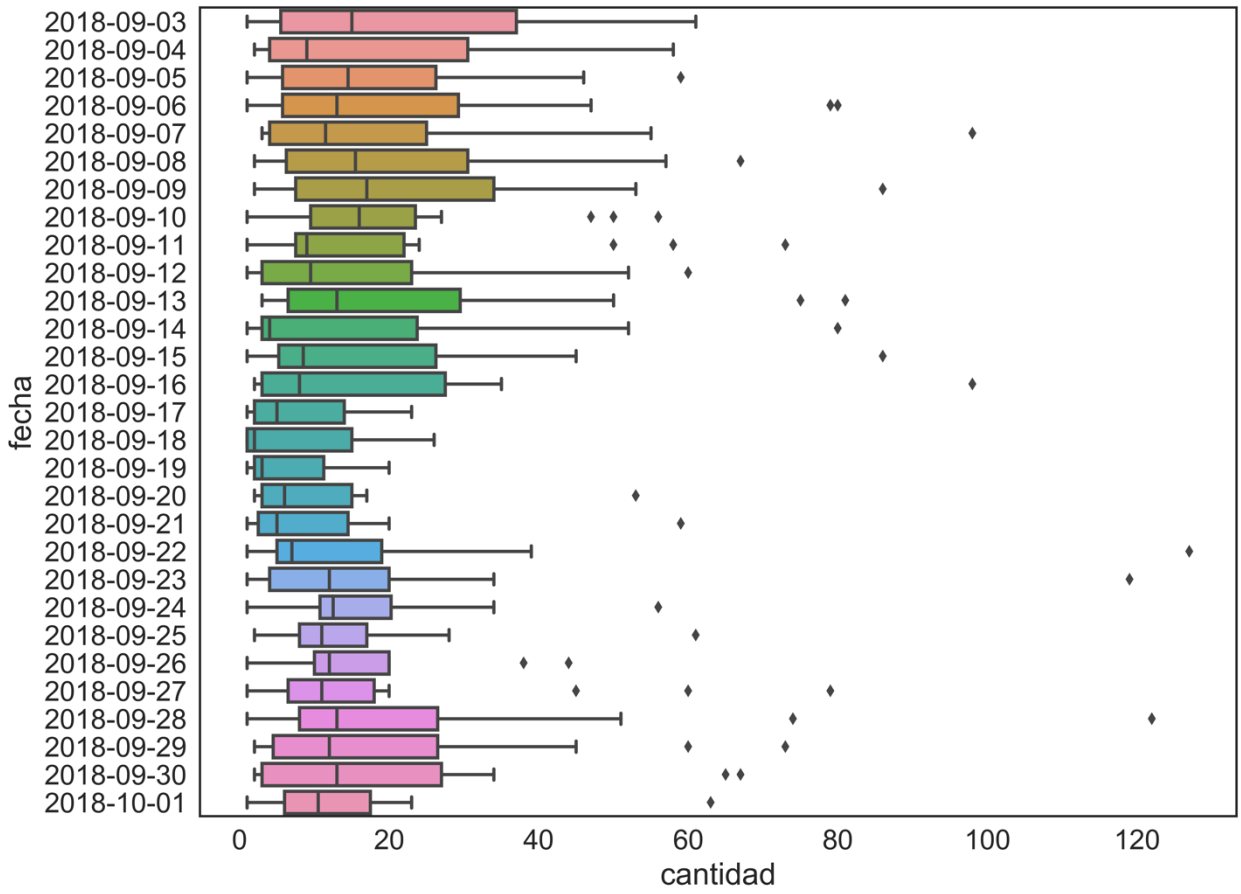

Imagen 36 - Cantidad de comerciales agrupados por etiqueta y fecha

Como resultado de la Imagen 36, es posible identificar que los datos se encuentran distribuidos durante el mes por diferentes cantidades de emisiones, agrupadas por etiquetas, y que para muchos días presenta una gran cantidad de *outliers*, los cuales se podrán explicar de mejor manera utilizando otro tipo de visualizaciones más adelante.

La Imagen 37 muestra un diagrama de cajas que agrupa el 100% de los datos por fecha y comercial.

La siguiente tabla muestra las estadísticas obtenidas para este análisis.

count	mean	std	min	25%	50%	75%	max
2535	2.9	3.1	1	1	2	4	39

Tabla 6 - Cantidad de comerciales por fecha

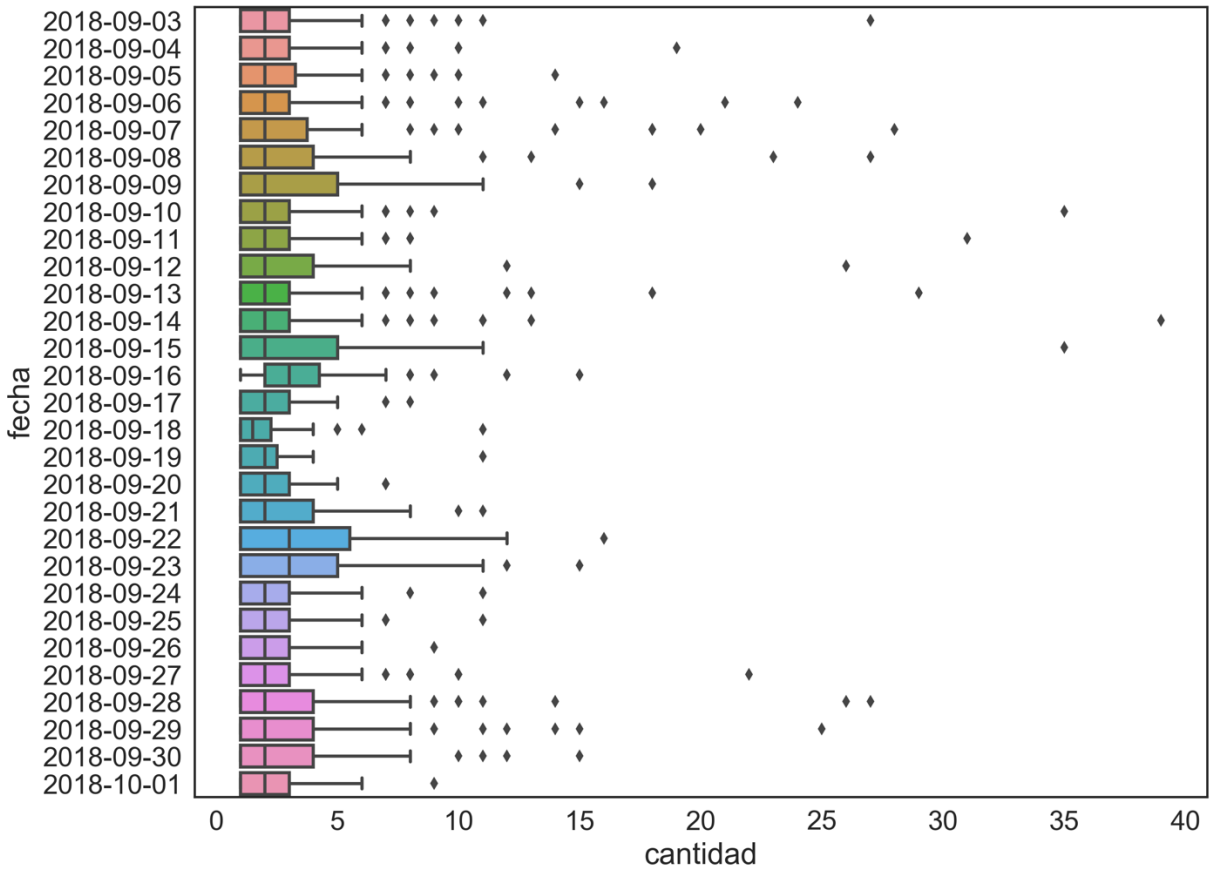

Imagen 37 - Cantidad de comerciales agrupados por fecha

Al agrupar los datos por cantidad de comerciales, es posible observar una distribución que presenta un patrón similar durante todas las semanas, registrando un mayor volumen de emisiones los días cercanos al fin de semana.

7.3. Mapas de Calor

Con el propósito de añadir una nueva dimensión al análisis de los datos, este tipo de diagrama resulta útil para comparar valores entre cantidad de emisiones, comerciales o etiquetas, y algún parámetro de tiempo.

A continuación se presentan diferentes visualizaciones de calor, con el objetivo de identificar aquellos comerciales que tienen una mayor presencia para ciertos días de la semana, como también, entender si existe alguna tendencia para el mes de las fiestas patrias Chilenas.

La siguiente tabla muestra las estadísticas obtenidas para este análisis.

count	mean	std	min	25%	50%	75%	max
78	15.6	7.0	10	11	12	18	39

Tabla 7 - Cantidad de comerciales por día

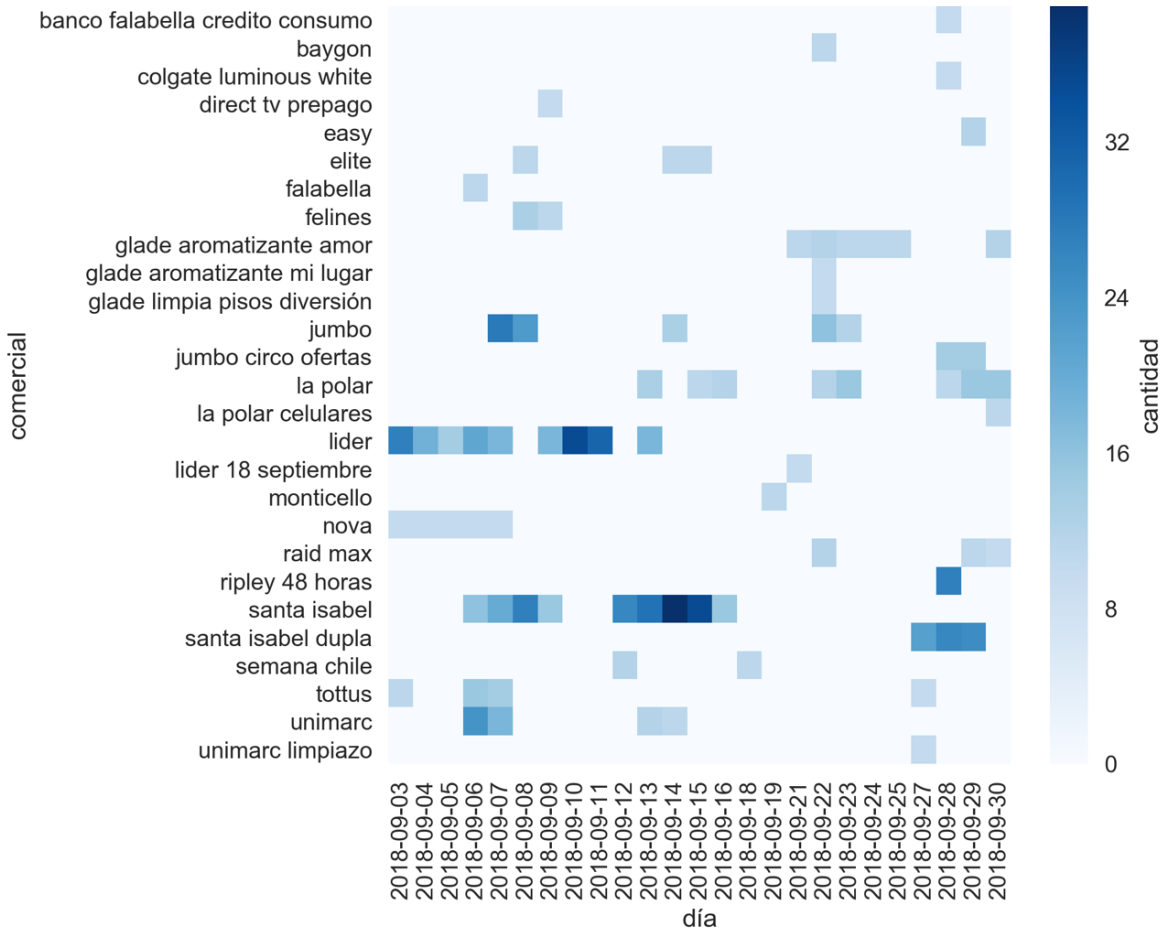

Imagen 38 - Cantidad de comerciales por día

Como resultado de la visualización anterior, se puede identificar que existe una gran cantidad de comerciales relacionados al rubro de los supermercados. Esta alta cantidad de emisiones se concentra en días consecutivos, previo a las fiestas patrias Chilenas.

Por el contrario, los festivos desde el día 17 al 23 presentan una clara disminución en la cantidad de emisiones por comercial. Sin embargo, resulta interesante identificar que existe una baja presencia en emisiones de *retail* durante el mes, previa celebración de fiestas patrias, y una tendencia al alza del mismo rubro, como también en los productos para el hogar en la última semana del mes.

Si bien las imágenes 36 y 37 identificaron una gran cantidad de *outliers* en los diferentes días del mes, la Imagen 38 permite identificar claramente aquellos comerciales que marcaron esta gran diferencia en los valores.

En la siguiente Imagen 39 se consideraron únicamente apariciones por etiqueta que registraron al menos 12 apariciones por fecha, entendiendo que esta cantidad comprende aquellas etiquetas que cubren la mayor parte del set de datos de pruebas, es representativa y no descarta datos relevantes para el análisis.

La siguiente tabla muestra las estadísticas obtenidas para este análisis.

count	mean	std	min	25%	50%	75%	max
198	32.7	23	12	16	24	45	127

Tabla 8 - Cantidad de comerciales por etiqueta

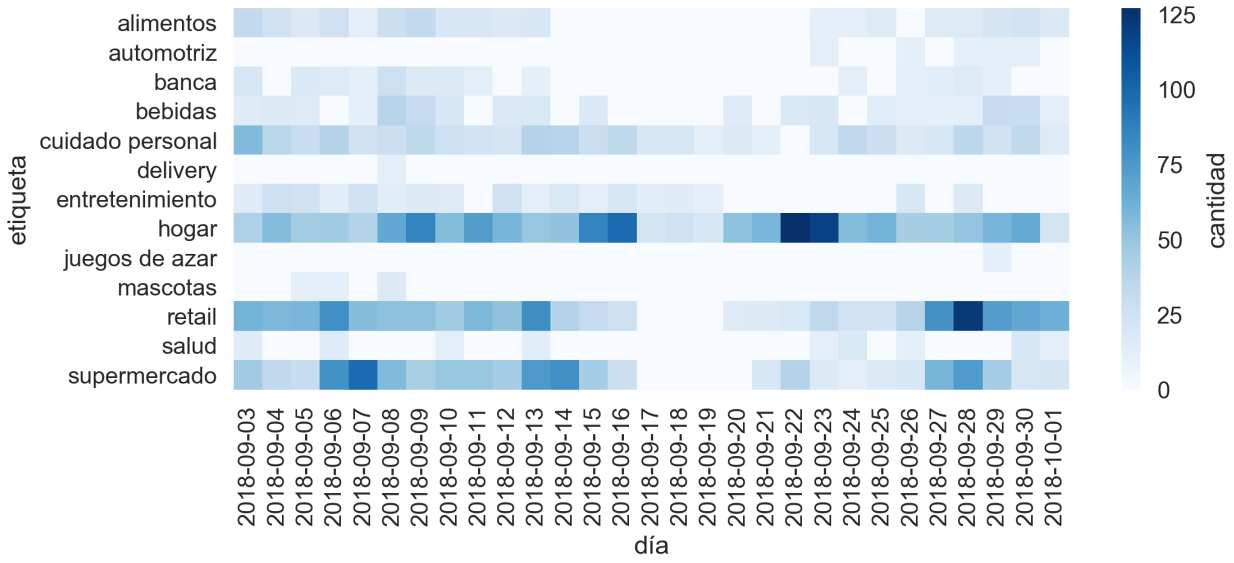

Imagen 39 - Cantidad de comerciales por etiqueta

Al visualizar los datos agrupados por etiquetas de la Imagen 39, es posible identificar con mayor facilidad que existe una gran presencia de comerciales relacionados a productos para el hogar. Sin embargo, durante los días festivos de Septiembre existen pocas emisiones de comerciales relacionados a supermercados y *retail*.

Un punto importante a destacar en esta visualización corresponde a la presencia casi completa de comerciales relacionados al cuidado personal, donde las farmacéuticas marcan presencia ofreciendo productos para el estómago, la gripe y venta de perfumes.

La Imagen 40 muestra la agrupación de comerciales por etiqueta para las diferentes horas del día en la cual son emitidos.

La siguiente tabla muestra las estadísticas obtenidas para este análisis.

count	mean	std	min	25%	50%	75%	max
267	28	40	1	5	14	32	312

Tabla 9 - Comerciales por hora del día según etiqueta

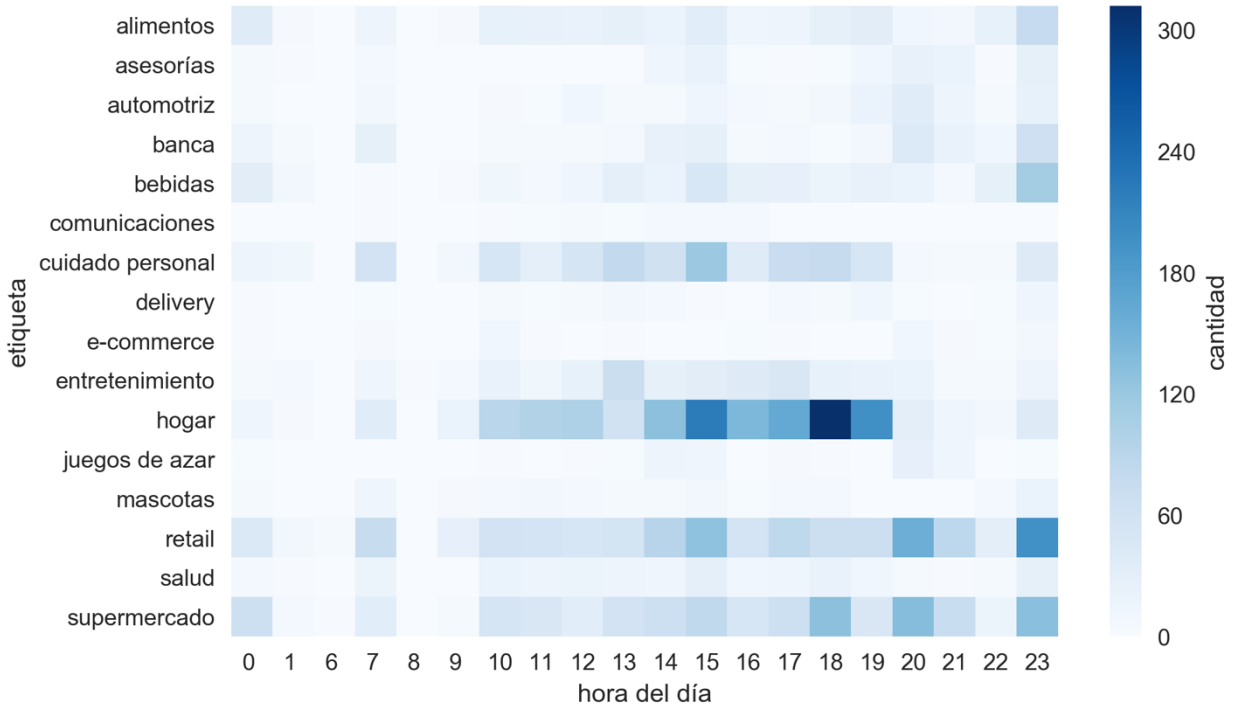

Imagen 40 - Comerciales por hora del día según etiqueta

En la Imagen 40 se puede identificar nuevamente a la etiqueta hogar registrando emisiones durante toda la tarde, con una tendencia al alza durante el término de la jornada laboral. En el horario para adultos, desde las 22 horas solo existe participación de etiquetas relacionadas al *retail*, supermercado, alimentos y bebidas.

La siguiente tabla muestra las estadísticas obtenidas para este análisis.

count	mean	std	min	25%	50%	75%	max
402	103.4	76.4	4	35	92.5	144.7	390

Tabla 10 - Cantidad de segundos por etiqueta según fecha

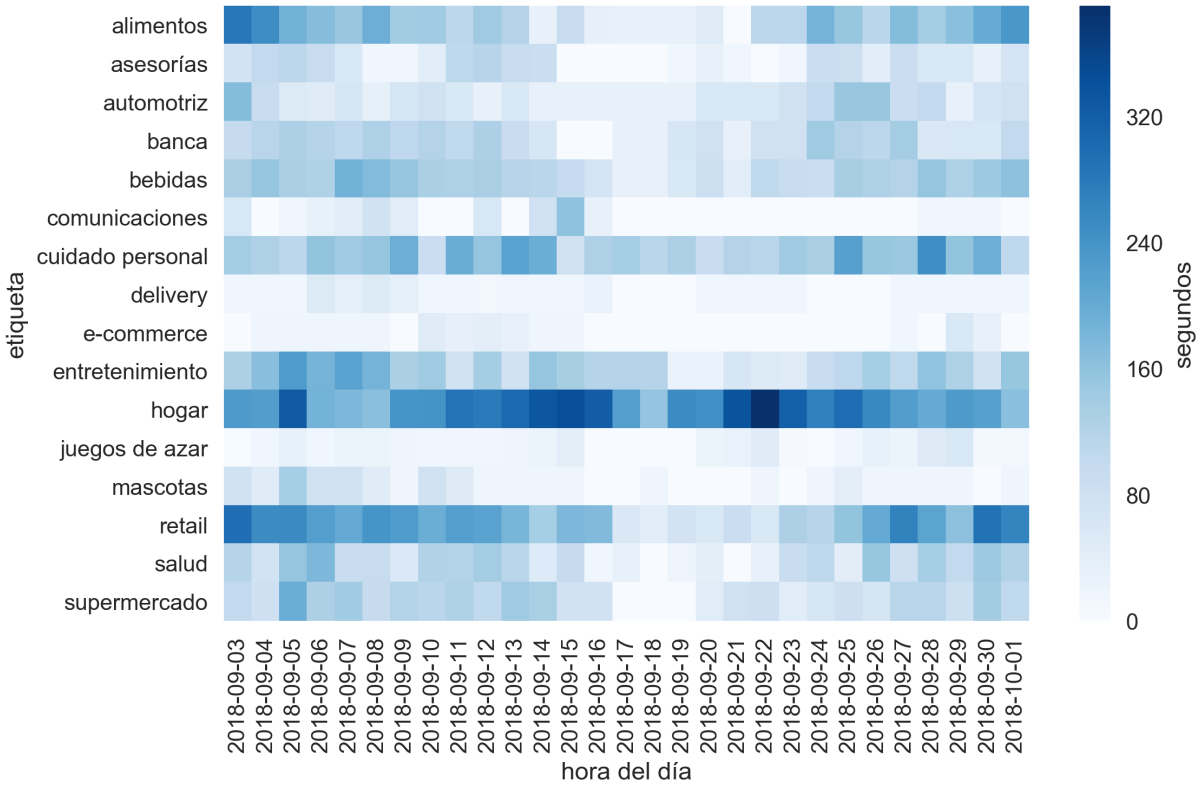

Imagen 41 - Cantidad de segundos por etiqueta según fecha

De acuerdo a la Imagen 41 es posible identificar, al igual que la Imagen 39, que la etiqueta hogar se mantiene presente durante la mayoría de los días analizados.

Por último, las siguientes visualizaciones corresponden a un análisis del mes completo agrupado por los días de la semana, con el propósito de identificar cuáles son las tendencias que se presentan durante las semanas del mes.

La siguiente tabla muestra las estadísticas obtenidas para este análisis.

count	mean	std	min	25%	50%	75%	max
122	61.4	42.1	1	31.25	54.5	84.75	212

Tabla 11 - Comerciales por hora según día de la semana

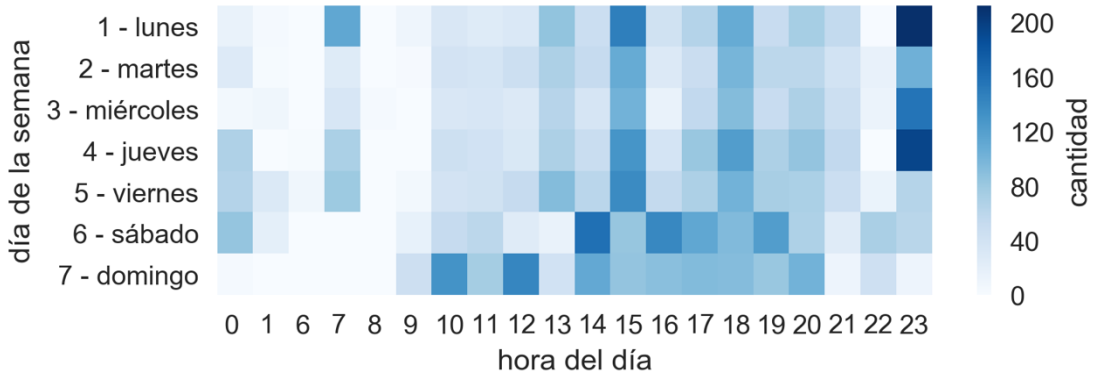

Imagen 42 - Comerciales por hora según día de la semana

En la Imagen 42, se puede identificar que el fin de semana registra un alza significativa de emisiones durante toda la jornada de la tarde. Sin embargo, la tendencia durante la semana parece estar ubicada solo a la hora de almuerzo, al regresar a casa, y luego a la hora de dormir.

Al analizar los datos por hora del día, resulta interesante entender cual es el comportamiento de las emisiones de comerciales para cada día de la semana e identificar si existe algúna tendencia para alguna etiqueta.

La siguiente tabla muestra las estadísticas obtenidas para este análisis.

count	mean	std	min	25%	50%	75%	max
111	67.49	75.74	1	16.5	36	91.5	368

Tabla 12 - Comerciales por etiqueta según día de la semana

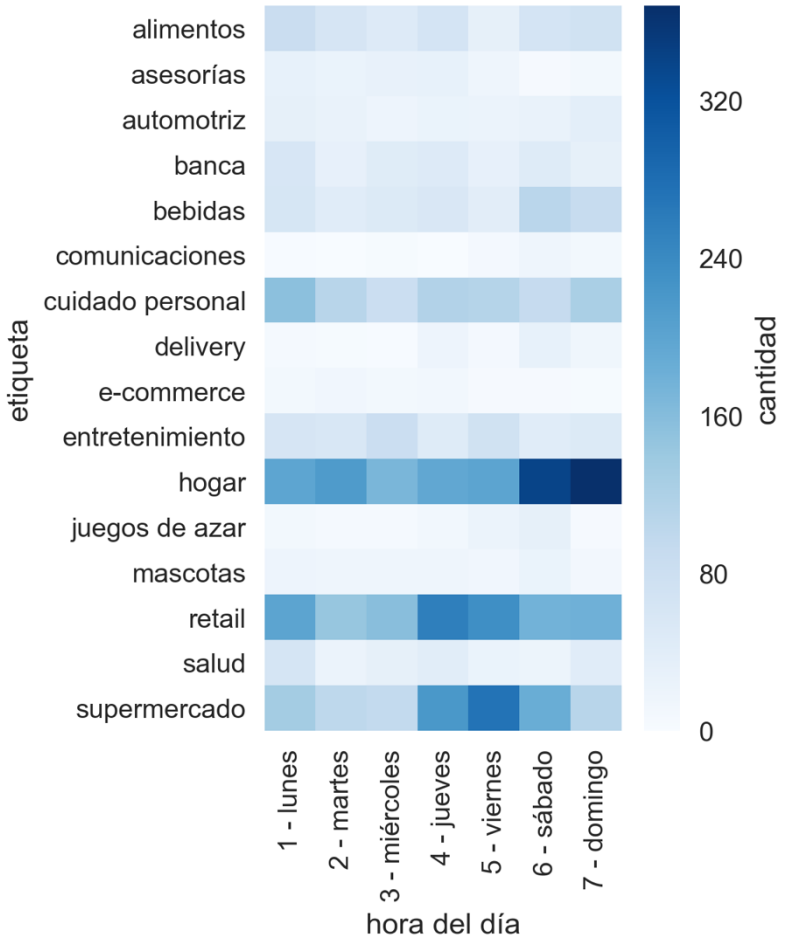

Imagen 43 - Comerciales por etiqueta según día de la semana

Con respecto a la Imagen 43, se puede concluir que nuevamente predominan los productos para el hogar, el *retail* se mantiene vigente durante la semana y los supermercados desde el día jueves comienzan a tener mayor participación en cantidad de emisiones.

8. Etapa de Evaluación

El siguiente punto trata en poder identificar cuan acertado es el algoritmo de detección, mediante la utilización de métricas mencionadas en la sección 2.9 como lo son *Precision* y *Recall*.

Aplicar esta medición debe considerar un espectro de diferentes experimentos, cuyo objetivo es entender cuáles son los parámetros de umbral que arrojan mejores, o peores, resultados en la detección de comerciales.

Teniendo en cuenta que este proyecto plantea objetivos relacionados a la implementación de algoritmos de detección, basados en diferentes descriptores, la etapa de evaluación considerará cuatro aspectos importantes en sus experimentos:

- Diferentes parámetros de umbral
- Detección utilizando descriptores de imagen
- Detección utilizando descriptores de audio
- Detección utilizando descriptores audiovisuales

El objetivo es mostrar la tasa de respuestas correctas que tiene el algoritmo de detección, para los diferentes parámetros de umbral en cada uno de los descriptores mencionados.

La etapa de evaluación se presentará utilizando visualizaciones de las curvas *Precision* y *Recall,* que permitirán visualizar cómo mejora, o empeora, la capacidad de detección, y concluirá con una visualización comparando los resultados MAP.

8.1. Métodos de detección

Desde el punto de vista de Recuperación de Información, determinar la cantidad de apariciones correctas en una colección de datos no es una tarea sencilla, como también, entender que tan preciso es el resultado recuperado.

La Recuperación de Información de un sistema puede conseguir resultados buenos y malos, siendo necesario evaluar cada uno de ellos cuando se realiza una consulta. Sin embargo, alcanzar un 100/100 para *Precision* y *Recall* en los resultados es muy difícil.

Una vez conseguidos los resultados, es importante identificar cual es la compensación *(trade-off)* de *Precision* y *Recall* para establecer un punto de equilibrio entre ambas medidas.

Con el objetivo de obtener la mejor compensación de resultados para este proyecto, se presentan diferentes métodos seleccionados de detección, a los cuales se les aplicaron cálculos de métricas para *Precision*, *Recall* y MAP.

Es importante considerar que, para cada método de detección listado en las siguientes secciones 8.1.2, 8.1.3 y 8.1.4, se utilizaron diferentes parámetros de umbral. El objetivo principal es obtener diferentes resultados para identificar la mejor compensación entre los resultados de *Precision* y *Recall*.

Como parte de la investigación de este proyecto, se identificó que al utilizar algoritmos de búsqueda aproximada, los resultados obtenidos inicialmente no fueron alentadores producto que muchas detecciones no estaban siendo recuperadas al realizar una consulta.

Las Imágenes 44 y 45 muestran cómo funcionan las detecciones y cómo se mitigó esta situación para poder recuperar los comerciales que inicialmente no estaban siendo considerados en la detección.

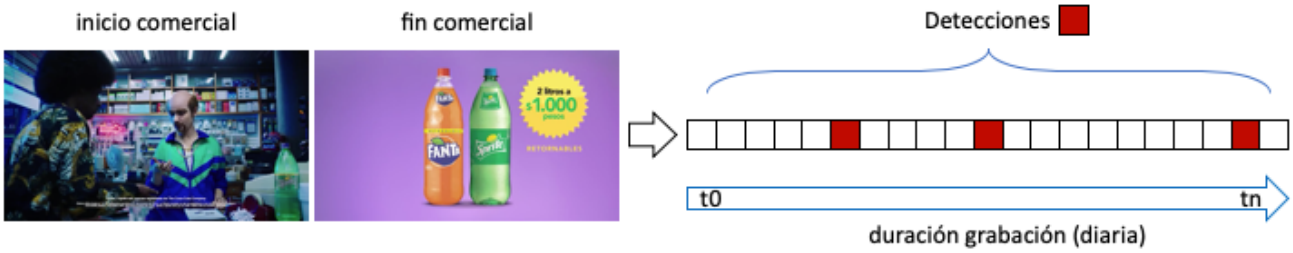

Imagen 44 - Detecciones consideradas

El algoritmo siempre considera el largo total del archivo multimedia que se quiera detectar, e intenta identificar, a lo largo de toda la grabación del día seleccionado, aquellas coincidencias que mayor similitud registren (cuadros rojos).

Muchos de estos comerciales no son detectados (cuadros grises) porque existen otras coincidencias que tienen mayor relevancia.

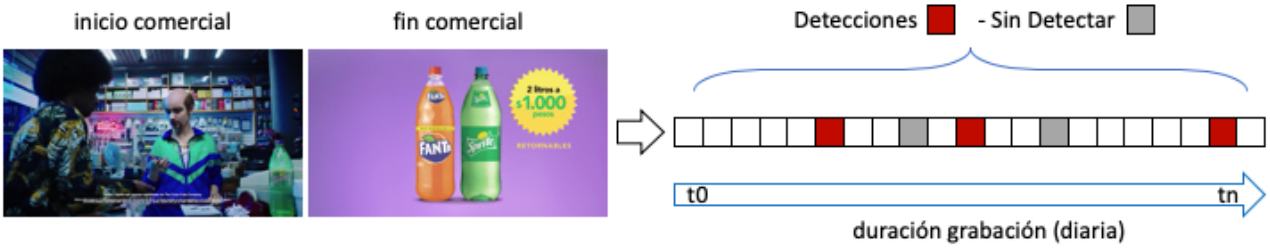

Imagen 45 - Detecciones no consideradas

Considerando que los cuadros grises también son relevantes, existe una necesidad de mitigar este escenario forzando al algoritmo para que considere aquellos casos con menor relevancia. Para lograr esta tarea, cada coincidencia detectada correctamente fue multiplicada por “0”, cuyo propósito busca anular el vector detectado.

Al volver a iterar por sobre la grabación completa del día, el vector ya detectado no es considerado y el algoritmo logra identificar los casos con menor relevancia. Luego de ser detectados, también son anulados bajo la misma operación (ver Imagen 46).

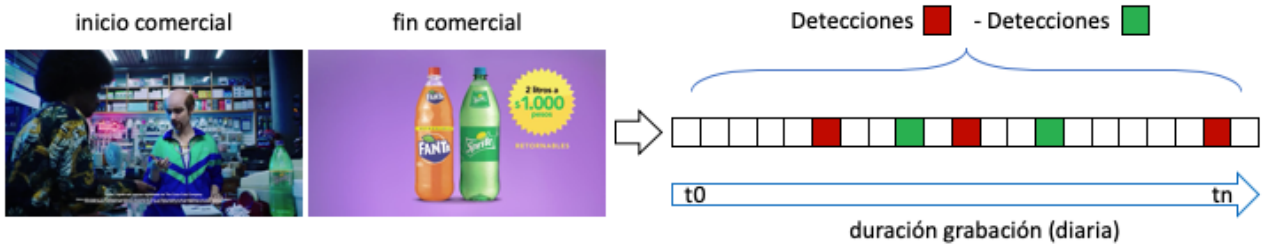

Imagen 46 - Detecciones consideradas con menor relevancia

Como resultado a lo expuesto anteriormente, las siguientes visualizaciones permiten identificar esta mejora con mayor facilidad al aplicar un umbral del 30%.

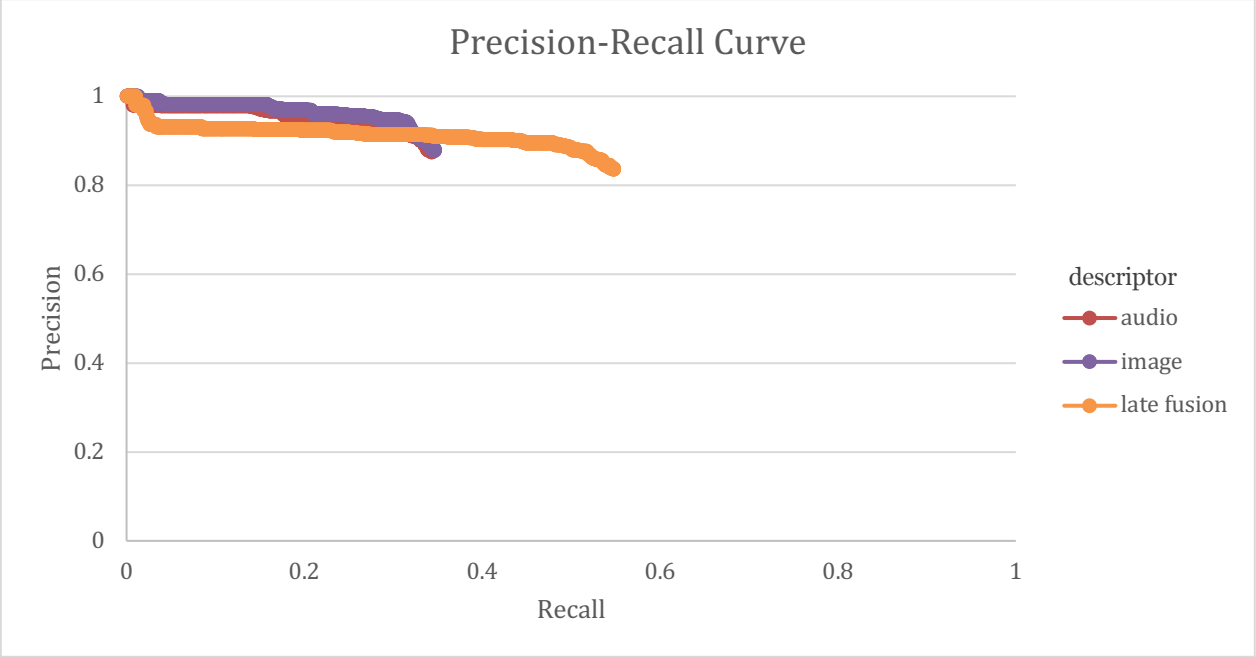

Imagen 47 - Detección sin anulación de vectores

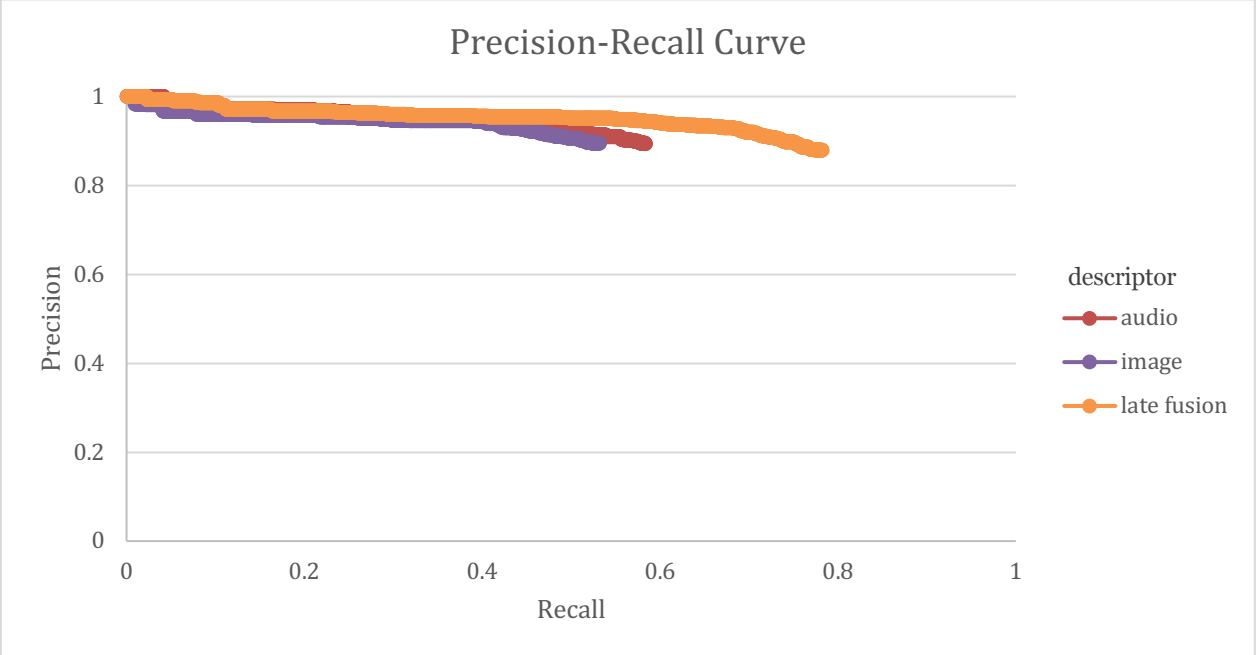

Imagen 48 - Detección con anulación de vectores

En comparación con la Imagen 47, al anular los vectores se puede alcanzar una mayor recuperación de los datos, tal como muestra la Imagen 48.

8.1.1. Parámetros de Umbral

En cuanto a la Recuperación de Información, es importante identificar cuáles son los resultados que serán considerados como válidos al ser recuperados.

Si un comercial tiene una duración de "n segundos", se espera que los resultados recuperados registren al menos una fracción de segundos de duración del comercial y debe responder ante la siguiente ecuación.

$$Detección = \sum frames\ recuperados \geq (Duración\ Comercial\ x\ Umbral)$$

El grado de relevancia resultante, para cada detección, es luego ordenado por la fracción de segundos de duración obtenida al realizar la detección de cada comercial.

Como parte de los experimentos, cuyo objetivo busca identificar cual es el mejor valor como parámetro de umbral, se realizarán pruebas utilizando los siguientes valores: 10%, 20%, 30%, 40%, 60%, 90%, 100%

8.1.2. Detección de Imagen

Con respecto a este tipo de descriptores, su composición está dada por el promedio de 3 *frames* por cada segundo y una dimensión de 10x10 píxeles, capturando la mayor cantidad de datos por cada lectura.

En la siguiente Imagen 49, se puede visualizar la capacidad de *Precision* y *Recall* de las detecciones utilizando únicamente el descriptor de imagen para los diferentes parámetros de umbral listados anteriormente.

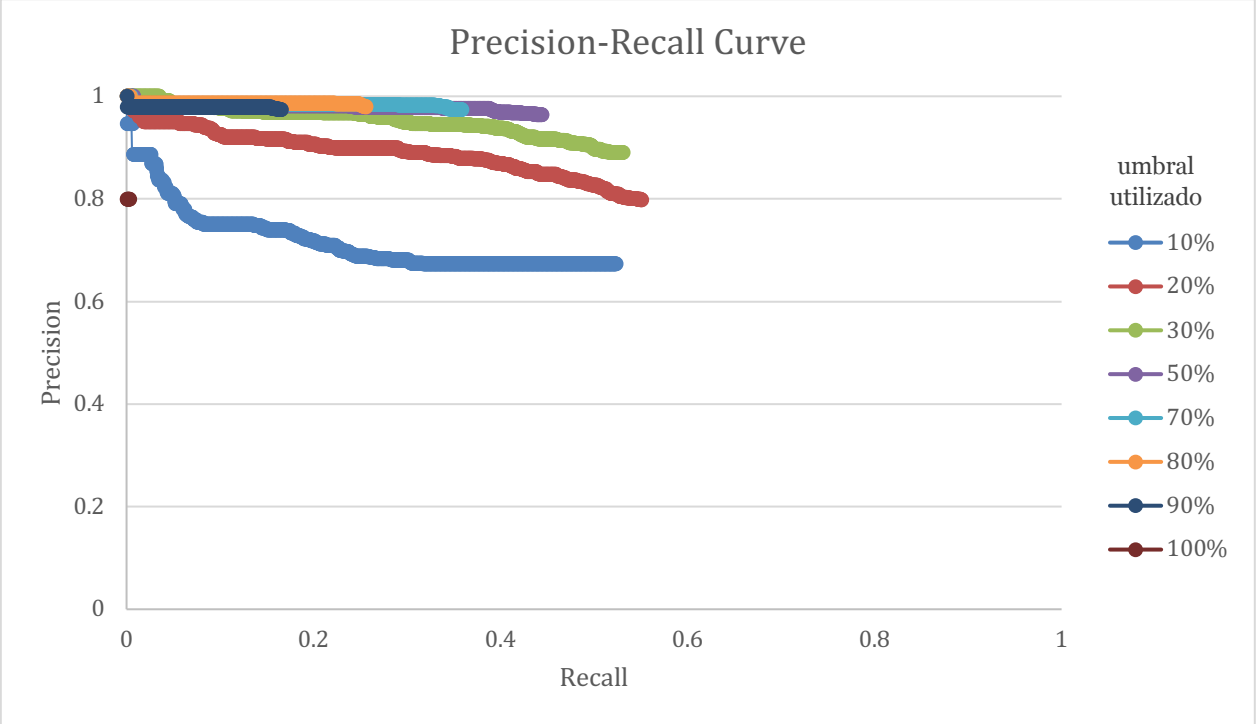

Imagen 49 - Detección utilizando descriptor de imagen

Como resultado de la Imagen 49, se puede visualizar que existen detecciones con *Recall* que alcanzan cerca del 55% del total de los datos.

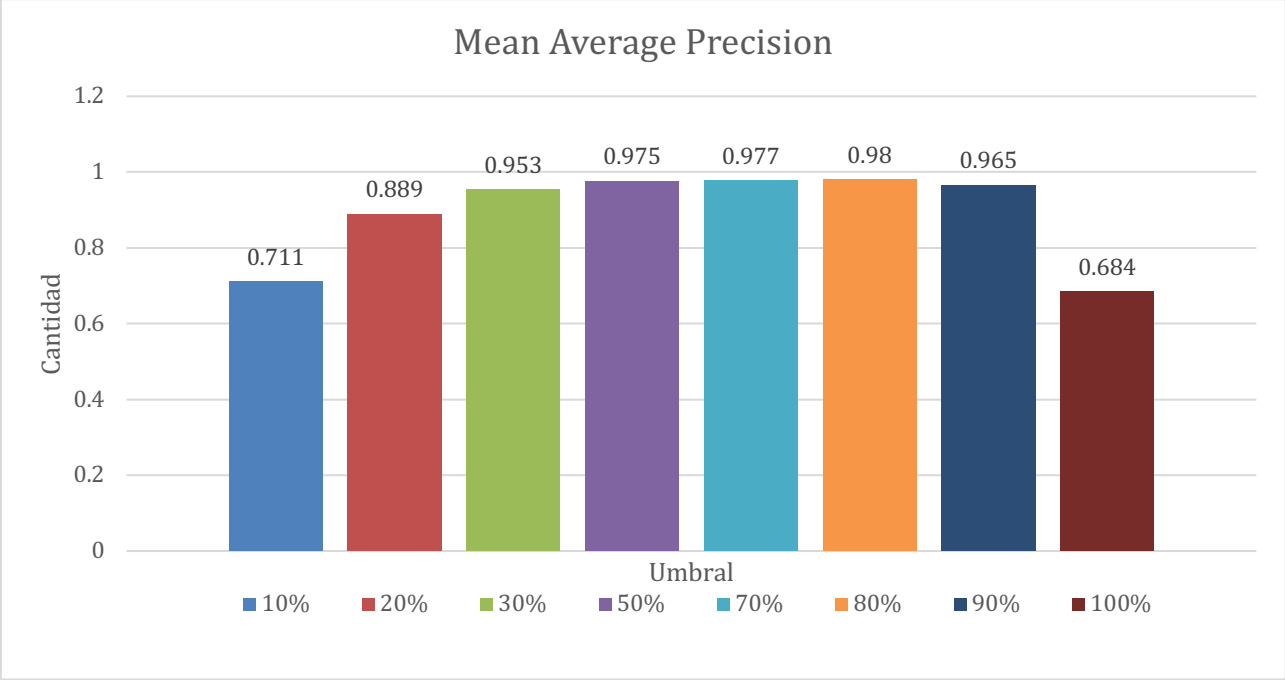

Imagen 50 - Resultados al detectar solo imágenes

Si bien *Recall* no supera el 55% de los datos recuperados, es importante destacar que *Precision* se mantuvo por sobre el 90% para las detecciones cuyo % de umbral fluctuó entre el 30% y 90%.

8.1.3. Detección de Audio

En cuanto a este tipo de detección, se realizó utilizando los descriptores MFCC, tal como fue descrito en la sección 6.2.

La Imagen 51 permite visualizar la capacidad de *Precision* y *Recall* de las detecciones, utilizando únicamente este tipo de descriptor para los diferentes parámetros de umbral listados anteriormente.

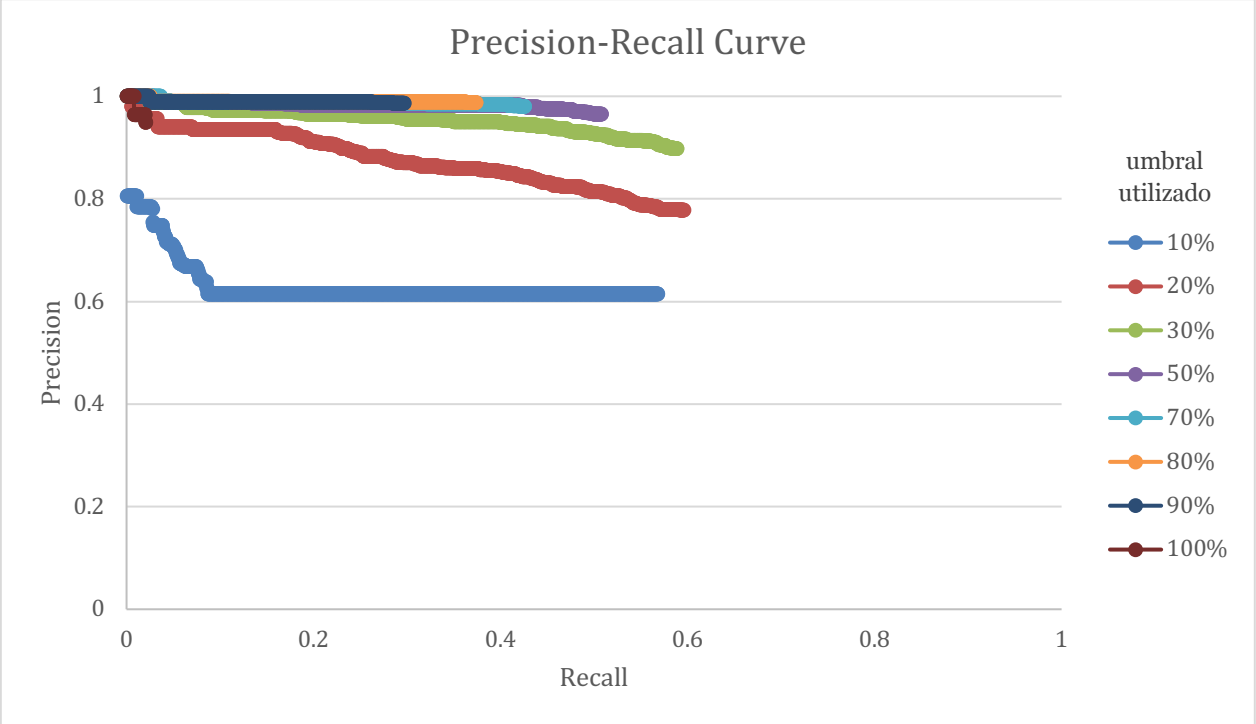

Imagen 51 - Detección utilizando descriptor de audio

Como resultado de la Imagen 51, se puede visualizar que se recuperaron comerciales utilizando el descriptor de audio. Sin embargo, con este descriptor el valor *Recall* solo alcanza cerca del 60% del total de los datos.

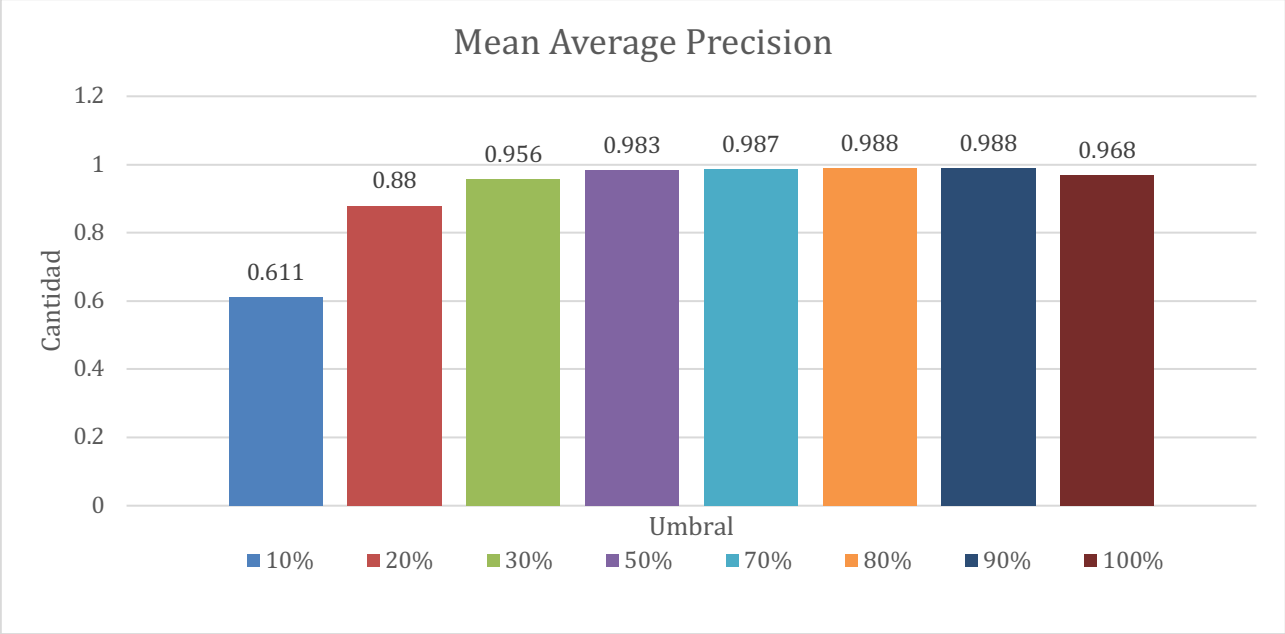

Imagen 52 - Resultados al detectar solo audio

A pesar que el valor *Recall* no supera el 60% de los datos recuperados, es importante destacar que el valor *Precision* se mantuvo por sobre el 90% para las detecciones, cuyo % de umbral fluctuó entre el 30% y 100%, marcando en este último % una diferencia considerable al ser comparado con el descriptor de imagen.

8.1.4. Detección Audiovisual *(Late Fusion)*

En relación a esta combinación de descriptores, se realizó utilizando los descriptores de audio y los descriptores de imagen mencionados anteriormente.

En la Imagen 53, se puede visualizar *Precision* y *Recall* utilizando únicamente este tipo de descriptor para los diferentes parámetros de umbral listados anteriormente.

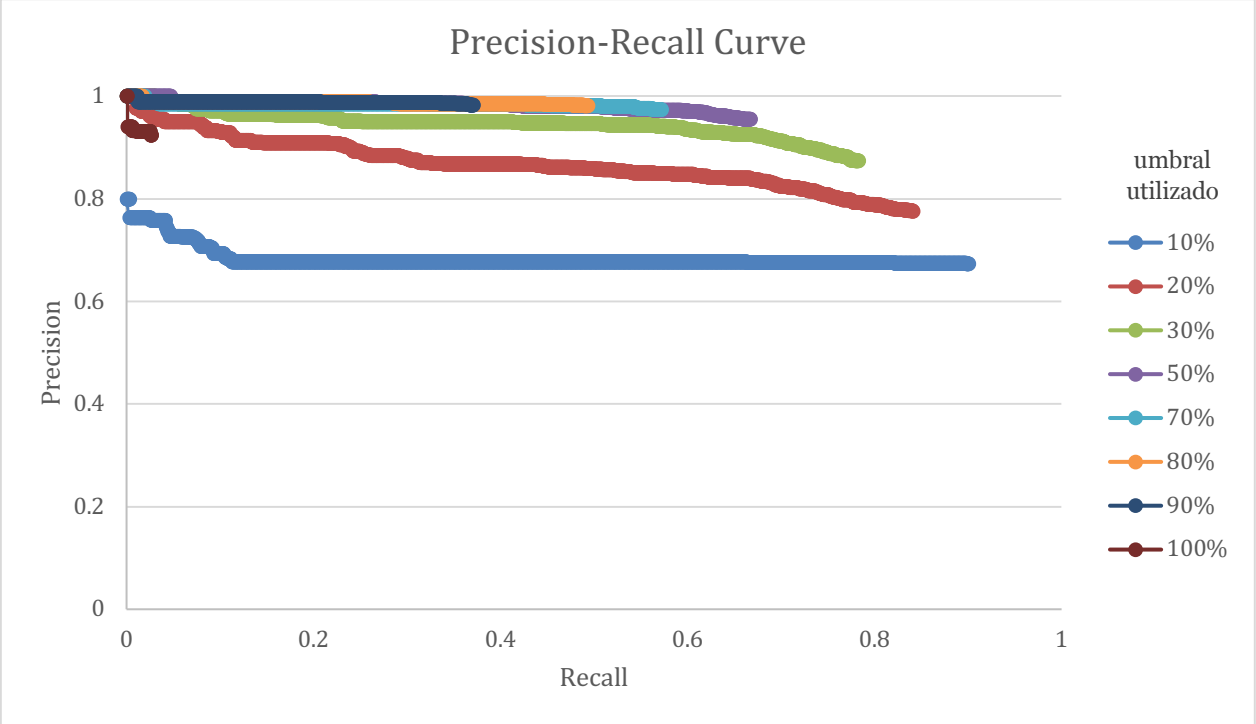

Imagen 53 - Detección utilizando descriptor de imagen y audio

Como resultado de la Imagen 53, se puede visualizar que se recuperaron comerciales utilizando el descriptor audiovisual. El valor *Recall* alcanza cerca del 90% del total de los datos cuando el umbral es del 10%. Sin embargo, el valor *Precision* se mantiene por debajo del 70%.

Por otro lado, umbrales entre el 20% y 30% alcanzan un valor *Recall* cercano al 80% y valor *Precision* por sobre el 75%.

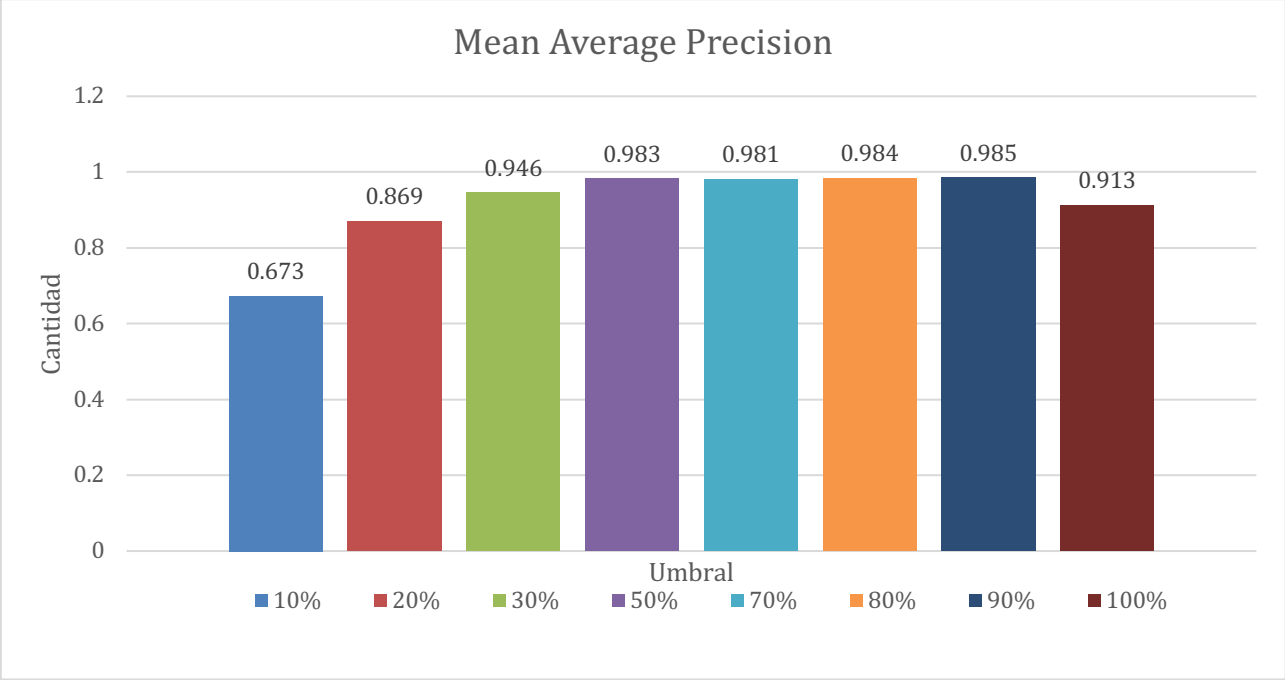

Imagen 54 - Resultados al detectar imagen y audio combinados

Como resultado de la Imagen 54, es importante destacar que *Precision* se mantuvo por sobre el 90% para las detecciones, cuyo % de umbral fluctuó entre el 30% y 100%.

A pesar de que el cálculo de MAP entre las imágenes 50, 52 y 54 no registra diferencias tan grandes, este último descriptor si presenta una mejora considerable para *Recall*.

9. Impacto de la Solución

Con el desarrollo de este proyecto, la cantidad de emisiones que ha tenido un comercial en el tiempo puede ser descubierta en pocos segundos utilizando como consulta archivos de video o pistas de audio. Esta característica facilita la búsqueda de contenido, realizando una búsqueda automática sobre la base de datos.

Poseer una herramienta que identifique automáticamente la cantidad de apariciones de comerciales, permitirá crear una base de datos para consultar los datos identificados, abriendo nuevas posibilidades de acceso a información relevante, pertinente, de acceso más sencillo y que puede ser utilizado para diferentes propósitos.

Según los datos analizados de publicidad, la televisión emite mayormente publicidad de productos para el hogar, supermercados, *retail* y dejando casi de lado a eventos culturales o de entretención.

Como muchos canales de televisión proporcionan información sobre sus parrillas televisivas para los diferentes meses del año, utilizar la información proporcionada por esta solución permitirá calcular el costo monetario que ha significado para los *Product Managers* el lanzamiento de algún comercial, como también, validar que la cantidad de emisiones contratadas se está cumpliendo con las diferentes entidades involucradas [R9].

10. Conclusiones

Como resultado de los experimentos expuestos en la sección 8, y dadas las hipótesis planteadas en este proyecto, se puede afirmar que es posible detectar secuencias duplicadas de contenido multimedia al analizar segmentos de video.

En particular, para los descriptores de imagen y audio, se puede determinar que por sí solos son capaces de identificar secuencias duplicadas.

A continuación se muestran otras conclusiones de este trabajo:

- De acuerdo a los experimentos realizados, es posible concluir que los descriptores de imagen y de audio, al operar individualmente, poseen una menor capacidad de recuperación de detecciones relevantes, producto que muchas veces el contenido puede registrar ruido que un descriptor no es capaz de resolver correctamente

- Al utilizar los descriptores por separado, se produce un efecto de baja *Precision* en los resultados. Es decir, en muchas ocasiones puede que la detección sea capaz de identificar correctamente gran parte del comercial, pero falla en una fracción mínima de tiempo para aquellos casos cuyo contenido cambió. Este suceso se presenta principalmente en comerciales de supermercados y *retail*, donde gran parte del comercial se mantiene, pero varía el tipo de producto que ofertan

- Conforme a los experimentos realizados con el descriptor *late fusion*, es posible determinar que su capacidad de detección mejora considerablemente en los resultados de recuperación, manteniendo un MAP similar a los otros dos descriptores utilizados

- Si se intersectan las detecciones entre ambos descriptores, se puede reducir considerablemente la tasa de falsos positivos para los comerciales que tengan características similares a los supermercados o *retail*, los cuales mantienen gran parte de su contenido y solo presentan diferencias en pequeñas fracciones de segundos

- Con respecto al método propuesto de anulación de vectores, la métrica *Recall* mejora considerablemente sin estropear *Precision*

Como consecuencia de la capacidad de detección que se demostró en la etapa de evaluación, considerar esta solución para los *Product Managers* no solo permitirá analizar los datos en un menor tiempo de respuesta, sino que también, reducirá la dependencia actual que los mantiene en espera por largas semanas y con una nula posibilidad de analizar los datos que generan información relevante.

A continuación se muestran algunas conclusiones que se pueden lograr al utilizar el detector de comerciales de este trabajo:

- De acuerdo a las visualizaciones expuestas en la etapa de exploración, fue posible identificar fácilmente que, previa fiestas patrias, los supermercados registraron una alta cantidad de emisiones que luego disminuyó considerablemente las siguientes semanas

- Los datos obtenidos en la etapa de análisis pueden servir para identificar cuál es la estrategia que las compañías pueden utilizar en el tiempo

- Con respecto al análisis de comerciales de Septiembre del año 2018, se puede concluir que existe una fuerte tendencia al alza en comerciales relacionados a productos para el hogar y de supermercados. Esto se puede manifestar por la fecha en la cual se realizó la exploración de datos, producto de las necesidades que se presentan para esas semanas de festejo

- Se puede identificar que durante los días de la semana, las compañías tienden a utilizar horas “punta” para emitir comerciales, los cuales al parecer, tienen como objetivo capturar la atención de las personas mientras están en instancias de descanso

Con la finalidad de recomendar un umbral, y de acuerdo a los experimentos realizados, se puede concluir que al aplicar un umbral del 30% es posible obtener un equilibrio entre los resultados de *Precision y Recall*, permitiendo recuperar la mayoría de los resultados con

una baja tasa de falsos positivos, lo que produce que no disminuya considerablemente *Precision*.

11. Trabajo Futuro

Como próximos pasos a seguir, el objetivo es ampliar las capacidades que esta solución posee actualmente, sumando nuevas fuentes de información, como reproductores multimedia en la web, y cuyo contenido resulta interesante analizar como parte de la experiencia que registran las personas al interactuar con este tipo de contenido publicitario en un entorno totalmente diferente a la televisión.

Uno de los objetivos principales es crear servicios para el acceso a la información recuperada, para que los interesados en trabajar con estos datos puedan generar estudios para beneficiar a la población, como también, extraer los datos para integrarlos con otras fuentes de información y complementar algún otro estudio relevante relacionado a esta materia.

Con respecto a las técnicas de detección de imagen, sería interesante evaluar que, si al extraer los descriptores de una red convolucional como *ResNet*, mejora la capacidad de recuperación, manteniendo un nivel de *Precision* similar al obtenido en los experimentos utilizando un descriptor de imagen 3fps 10x10 píxeles.

Por último, añadir capacidades de búsqueda, mediante la utilización de contextos, corresponde a una de principales metas futuras que permitirán otorgar al usuario la capacidad de realizar búsquedas por acciones propias que se realicen dentro del comercial.

En muchas ocasiones las personas son capaces de recordar una acción, característica, u otra descripción que no corresponde al nombre oficial de lo que vio o escuchó en algún momento del tiempo. Sin embargo, describir una acción podría resultar una pieza fundamental de búsqueda para recuperar contenido relacionado y proporcionar resultados similares a lo descrito.

Resultaría interesante trabajar el contenido multimedia realizando búsquedas de comerciales utilizando descripciones relacionadas a un contexto. Por ejemplo, “mujer corriendo en el parque”, “hombre cocinando”, “familia paseando en la nieve”, podrían permitir acceder a las propuestas de comerciales que utilizan las entidades en diferentes países para motivar a las personas en adquirir una vida saludable, con mayor deporte, propuestas de entretención en familia, entre otros, lo que podría impactar directamente en el diseño de nuevas estrategias publicitarias.

12. Bibliografía

[1] Stephen Moore, Deep Learning for Computer Vision, Expert Techniques, 2018

[2] Roberto Raieli, Multimedia Information Retrieval, Theory and Techniques, 2013

[3] Richard Szeliski, Computer Vision, Algorithms and Applications, 2011

[4] Henk Blanken, Arjen P. De Vries, Henk E. Blok, Ling Feng, Multimedia Retrieval, 2007

[5] François Chollet, Deep Learning with Python, 2018

[6] Meinard Müller, Fundamentals of Music Processing. Audio, Analysis, Algorithms, Application, 2015

[7] Rafael C. González, Richard E. Woods, Digital Image Processing, Second Edition, 2002

[8] Peter Knees, Markus Schedl, Music Similarity and Retrieval, 2016

[9] Hanan Samet, Foundations of Multidimensional and Metric Data Structures, 2006

[10] Ricardo Baeza-Yates, Berthier Riveiro-Neto, Modern Information Retrieval, 1999

[11] Alan C. Bovik, Handbook of Image Video Processing, 2000

13. Referencias

[R1] https://www.ukessays.com/essays/information-technology/role-of-multimedia-in-todays-society-information-technology-essay.php, Última visita: Agosto 2019

[R2] https://es.wikipedia.org/wiki/Comercial_de_televisión, Última visita: Agosto 2019

[R3] https://www.zenithmedia.com/26-of-media-consumption-will-be-mobile-in-2019 Última visita: Julio 2018

[R4] https://www.thinkbox.tv/News-and-opinion/Newsroom/Why-TV-remains-the-worlds-most-effective-advertising Última visita: Julio 2018

[R5] https://www.statista.com/statistics/217134/total-advertisement-revenue-of-super-bowls Última visita: Julio 2018

[R6] https://www.recode.net/2017/12/4/16733460/2017-digital-ad-spend-advertising-beat-tv Última visita: Julio 2018

[R7] https://www.engadget.com/2010/08/02/how-to-tweak-showanalyzer-to-100-percent-commercial-detection-ac Última visita: Julio 2018

[R8] https://www.nidcd.nih.gov/health/age-related-hearing-loss Última visita: Mayo 2019

[R9] https://www.tvn.cl/comercial/parrillas-y-tarifas Última visita: Mayo 2019

[R10] https://www.researchgate.net/publication/267507430_Content-Based_Video_Copy_Detection Última visita: Mayo 2019

[R11] https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html Última visita: Agosto 2019

[R12] http://www.comskip.org/ Última visita: Septiembre 2019

[R13] https://showanalyzer.software.informer.com/ Última visita: Septiembre 2019

[R14] https://mpeg.chiariglione.org/standards/mpeg-7 Última visita: Octubre 2019

[R15] https://opencv.org/about/ Última visita: Diciembre 2019

[R16] http://conference.scipy.org/proceedings/scipy2015/pdfs/brian_mcfee.pdf Última visita: Diciembre 2019

[IMG 1] https://ai.stanford.edu/~syyeung/cvweb/tutorial1.html Última visita: Septiembre 2019

[IMG 2] Alan C. Bovik, The Essential Guide to Image Processing, 2009, Pág. 9

[IMG 3] Richard Szeliski, Computer Vision, Algorithms and Applications, 2011, Pág. 98

[IMG 4] Gary Bradski, Adrian Kaehler, Learning OpenCV-O'Reilly, 2008, Pág. 148

[IMG 5] Alan C. Bovik, The Essential Guide to Image Processing, 2009, Pág. 16

[IMG 7] Müller, Meinard, Fundamentals of Music Processing, Audio, Analysis, Algorithms, Applications, 2015, Pág. 21

[IMG 9] Peter Knees, Markus Schedl, Music Similarity and Retrieval, An Introduction to Audio- and Web-based Strategies, 2016, Pág. 37

[IMG 10] Müller, Meinard, Fundamentals of Music Processing, Audio, Analysis, Algorithms, Applications, 2015, Pág. 20

[IMG 11] Peter Knees, Markus Schedl, Music Similarity and Retrieval, An Introduction to Audio- and Web-based Strategies, 2016, Pág. 34

[IMG 12] Peter Knees, Markus Schedl, Music Similarity and Retrieval, An Introduction to Audio- and Web-based Strategies, 2016, Pág. 40

[IMG 13] Peter Knees, Markus Schedl, Music Similarity and Retrieval, An Introduction to Audio- and Web-based Strategies, 2016, Pág. 54

[IMG 14] https://www.researchgate.net/figure/Extraction-process-of-MFCC-coefficients_fig1_284102947 Última visita: Septiembre 2019

[IMG 15] Peter Knees, Markus Schedl, Music Similarity and Retrieval, An Introduction to Audio- and Web-based Strategies, 2016, Pág. 56

[IMG 16] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, 2016, Pág. 150

[IMG 17] Alan C. Bovik, The Essential Guide to Image Processing, 2009, Pág. 7

[IMG 22] Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, Michal Batko, Similarity Search, The Metric Space Approach, 2006, Pág. 91

[IMG 23] Hanan Samet, Foundations of Multidimensional and Metric Data Structures, 2006, Pág. 50

[IMG 24] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval, 1999, Pág. 75

[IMG 25] https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html Última visita: Agosto 2019