UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

TRABAJO DE FIN DE MÁSTER

Indexación automática de vídeos de noticiarios mediante extracción de características visuales globales.

Máster Universitario en Ingeniería de Telecomunicación

Autor: Sara Cerro Pardo
Tutor: José M. Martínez

Junio 2016

Indexación automática de vídeos de noticiarios mediante extracción de características visuales globales.

AUTOR: Sara Cerro Pardo
TUTOR: José María Martínez Sánchez

Grupo VPULab
Dpto. de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid

Junio 2016

Agradecimientos

En primer lugar, quería dar las gracias a mi tutor, Chema, por haberme ofrecido su asesoramiento en todo momento y haber tenido siempre un hueco disponible para cualquier cosa que pudiera necesitar.

A todas las personas del laboratorio, por su disposición a ayudar y por crear un buen ambiente que hace más amenas las horas de trabajo.

Tras estos años de Grado y Máster, tengo que agradecer a todo el profesorado de la Escuela su labor en formarnos e intentar sacar lo mejor de nosotros, con especial mención a aquellos que se han involucrado más, mostrando su cercanía y poniendo todo su empeño para que hoy estemos aquí.

A todos mis compañeros durante estos años, que al final nos hemos convertido en una pequeña familia y en grandes amigos; porque sin ellos estoy convencida de que no hubiera sido lo mismo.

Sobre todo, gracias a mi equipo de remo, por todos los momentos compartidos y los muchos que aún nos quedan.

A mis amigos de toda la vida, que siempre están y me lo han demostrado cuando he necesitado un empujón. A Patri, porque la mencionaría en todas las memorias que escribiera durante mi vida, por demostrarme lo que la amistad significa y recordármelo cada día.

A todos aquellos que, de algún modo u otro, han formado una parte importante de mi vida durante estos años y han puesto su granito de arena para que hoy esté aquí.

Por último, quería dar las gracias a mis padres, Manolo y Pepa, y a mi hermana Isabel; porque la familia no se escoge pero yo no he podido tener más suerte. Gracias por vuestra paciencia, la confianza, la fuerza y el amor que me dais día tras día. Sin vosotros hoy no sería quien soy.

Sara Cerro Pardo
Junio 2016

Resumen

La motivación principal de este Trabajo de Fin de Máster ha sido desarrollar un sistema de indexación automática de vídeo, concretamente de programas informativos o noticiarios, destinado principalmente a agilizar el proceso de búsqueda de información de interés en vídeos almacenados en bases de datos, cuya dimensión se ha visto disparada en los últimos años con la gran cantidad de información en formato digital y accesible a través de Internet.

Dado este elevado crecimiento en la información de vídeo disponible y dados los múltiples usos y aplicaciones en los que un sistema de este tipo tiene cabida, se trata de un campo de estudio en el que se investiga activamente.

Por ello, el sistema propuesto de indexación de vídeo está basado en un amplio estado del arte, con numerosas implementaciones y diferentes técnicas usadas para dicho fin.

El objetivo primordial perseguido durante la duración de este trabajo ha sido crear un algoritmo automático y generalista de indexación por medio de la extracción, análisis e indexado de características visuales globales, así como de la combinación de las mismas, de lo que puede llegar a obtenerse información muy relevante.

Se ha buscado conseguir unos resultados razonables en las pruebas realizadas con los vídeos bajo estudio, aportando diferentes enfoques con los que establecer conclusiones.

De igual modo, se han comentado los problemas inherentes a este tipo de implementaciones y, más en concreto, aquellas dificultades que han surgido durante el tiempo de trabajo.

Finalmente, se han asentado las bases futuras en las que centrar los esfuerzos para corregir dichos problemas y crear, de esta manera, un sistema más robusto y eficiente.

Palabras clave

Indexación automática de vídeo de noticiarios, recuperación de vídeo, sistemas de indexación basados en contenido, segmentación en unidades homogéneas, extracción de características.

Abstract

This Master Thesis has been motivated by the need to develop an automated video indexing system, specifically for news videos, mainly aimed at speeding up the process of finding relevant information in stored videos databases, whose dimension has been highly increased in recent years due to the quantity of raw video data in digital format and accessible by the Internet.

Given the constant growth in available video information and given the multiple uses and approaches in which a system like this could be applied, it is currently being actively investigated in this field.

This system based on video-indexing has a large state of the art, with numerous implementations and different techniques used for this purpose.

The main target during the development of this Project has been to create an automated and general indexing algorithm by visual and global feature extraction and analysis, testing the algorithm with datasets of news videos, obtaining relevant information and providing some approaches to set conclusions.

In the same way, the problems attached to this type of implementations have been explained and, in particular, the difficulties that have appeared throughout this work period.

Finally, it has been established the main basis for future studies on how to correct these problems and create at the same time a more robust and efficient algorithm.

Keywords

News video automated indexing, information retrieval, content-based indexing, semantic video segmentation, systems, feature extraction.

Índice de contenido

Índice de figuras

Índice de tablas

Capítulo 1. Introducción

1.1 Motivación

Hasta hace un par de décadas, las bases de datos de vídeo eran relativamente pequeñas, por lo que la indexación, análisis y recuperación de los vídeos se basaba en la anotación de palabras clave de forma manual.
Hoy en día, sin embargo, dichas bases de datos han visto disparado su crecimiento [1], dada la cantidad ingente de vídeos en formato digital y accesible a través de internet.

En vista de este progreso y su tendencia a incrementarse, se hace necesaria la posibilidad de buscar eficientemente en repositorios de vídeo [1][2][3][4]. Para que dichas búsquedas resulten efectivas es necesario que existan índices o etiquetas de los vídeos en función de su contenido y que, además, se generen de la manera más automática posible, minimizando al máximo la intervención humana.

Las operaciones de búsqueda y recuperación de vídeos tienen un amplio rango de aplicaciones, entre las cuales podemos encontrar distintos ejemplos:

- Navegación rápida a través de carpetas de vídeo.
- Gestión inteligente de los videos de la web (búsqueda de vídeo útil y rastreo de vídeo perjudiciales).
- Análisis del comercio electrónico visual o de eventos de prensa (sobre las tendencias de interés seleccionadas por los usuarios o las correlaciones entre los anuncios y sus efectos)
- Ámbito de la enseñanza a distancia y/o museos digitales.
- Video vigilancia.

Es esta amplia gama de aplicaciones [1], en las que la indexación de vídeo juega un papel fundamental, lo que ha motivado el interés de investigadores en este campo alrededor del mundo.
No obstante, se trata de una tarea que no es trivial, dado el variable nivel de detalle por el que pueden indexarse distintos tipos de vídeo, los cuales poseen numerosas características diferentes y que pueden disponer de análisis muy dispares en función del contexto en que se pretenda darle uso.

1.2 Objetivos

Con este proyecto se pretende crear un algoritmo automático y generalista de indexación de vídeos por medio de la extracción, análisis e indexado de características visuales globales, así como de la combinación de las mismas, de lo que puede llegar a obtenerse información muy relevante.

Los objetivos principales que se persiguen durante el trabajo son los siguientes:

- Realizar un estudio del estado del arte sobre la evolución de las bases de datos de vídeo y los distintos métodos y algoritmos utilizados de cara a la **indexación de vídeo de forma masiva para su posterior recuperación**.
- Proponer un algoritmo capaz de etiquetar vídeos de **forma automática** basándose en su contenido, lo cual **reducirá notablemente el tiempo empleado en recuperar material multimedia,** que podrá disponer de un **amplio uso en diferentes contextos.**
- Estudiadas y analizadas las posibles vías para darle solución, se procederá a **implementar el algoritmo propuesto** en lenguaje C++ a través de OpenCV y sus librerías.

Finalmente, se realizarán pruebas con vídeos de noticiarios y se **medirá su rendimiento**, proponiendo posibles mejores y **sirviendo de base a trabajos futuros**.

1.3 Organización de la memoria

La memoria se organizará en seis capítulos diferentes, estando dedicado cada uno de ellos a un bloque específico:

- **Capítulo 1:** Introducción al trabajo y organización del mismo.
- **Capítulo 2:** Estado del arte general sobre el tema a tratar.
- **Capítulo 3:** Estado del arte específico sobre el trabajo a llevar a cabo.
- **Capítulo 4:** Diseño y desarrollo del sistema propuesto.
- **Capítulo 5:** Ejecución de pruebas y evaluación de los resultados obtenidos.
- **Capítulo 6:** Líneas de trabajo futuras y conclusiones finales.

Capítulo 2. Estado del arte

2.1 Introducción

Aunque cada vídeo posee unas características concretas que lo hacen único, a partir de las cuales se pueden extraer etiquetas distintivas que lo diferencian de otros, de manera general un vídeo posee las siguientes propiedades [7]:

1. Contenido mucho más rico que una imagen individual.
2. Gran cantidad de datos en bruto (*raw data*).
3. Muy poca estructura a priori.

Estas propiedades hacen a la indexación (*indexing*) y a la recuperación (*retrieval*) de vídeo una tarea arduamente complicada.

Un vídeo puede ser representado de las siguientes maneras [5]:

- Por su **contenido de audio**. Por ejemplo, su música de fondo o música ambiente, las diferentes voces presentes en él o los sonidos producidos por objetos tales como el claxon de un coche o el timbre de un teléfono.
- Por **contenido de texto**. Ejemplos de ello son los subtítulos en una película, los titulares de noticias de un informativo, los nombres de los jugadores o los resultados en un evento deportivo o el nombre de un producto en anuncios publicitarios.
- Por su **contenido visual.** Suele ser el contenido más abundante y el que caracteriza en gran parte a la idea o al propósito del vídeo. Paisajes, objetos o personas en movimiento son sólo alguno de los múltiples contenidos visuales que pueden extraerse de un vídeo.

Todas estas características identifican de manera unívoca a cada vídeo y, del mismo modo, el número de estas características aumenta a medida que se aumenta también la granularidad en la que se analiza dicho vídeo. En definitiva, la información contenida en un vídeo es inmensa y se encuentra, a priori, desestructurada. Aun así, es posible distinguir ciertas partes identificativas para cada tipo o género en el que pueden clasificarse los vídeos. Por ejemplo, en una película son fácilmente reconocibles las etapas de la introducción, trama y desenlace; en un informativo se pueden diferenciar con exactitud distintas partes (introducción de la noticia, reportaje, cabeceras, etc.). Sin embargo, dependiendo del **nivel de detalle** que se pretenda obtener, así como del **fin que se persiga** con la extracción de información, se pueden generar un gran número de características distintas y asociarlas al vídeo bajo análisis.

El problema reside en que, actualmente, supone un gran reto para los motores de búsqueda filtrar este tipo de características inherentes en un vídeo si previamente no han sido anotadas de forma rigurosa, lo cual constituye una tarea realmente complicada dado el **enorme volumen de vídeos almacenados en bases de datos disponible a través de Internet**, donde se vuelve claramente imposible realizar la anotación de manera manual.

Como resultado a esta problemática, surgen novedosas **técnicas y metodologías** que son capaces de manipular grandes cantidades de información de acuerdo a su contenido. Del mismo modo, se han creado **algoritmos** muy variados que se centran en diferentes aspectos para que sea posible la eficiente indexación y anotación de los mismos.
Esta información puede usarse en muchos campos de aplicación como en librerías digitales, en el ámbito educativo, las comunicaciones y el entretenimiento.

2.2 Etapas

En un proceso común de indexación y recuperación de vídeo es posible distinguir de forma concreta diferentes etapas [2][4], las cuales pueden observarse en la Figura 2.1.
En ella se observan dos procedimientos en paralelo de una operación usual a la hora de buscar un vídeo determinado.

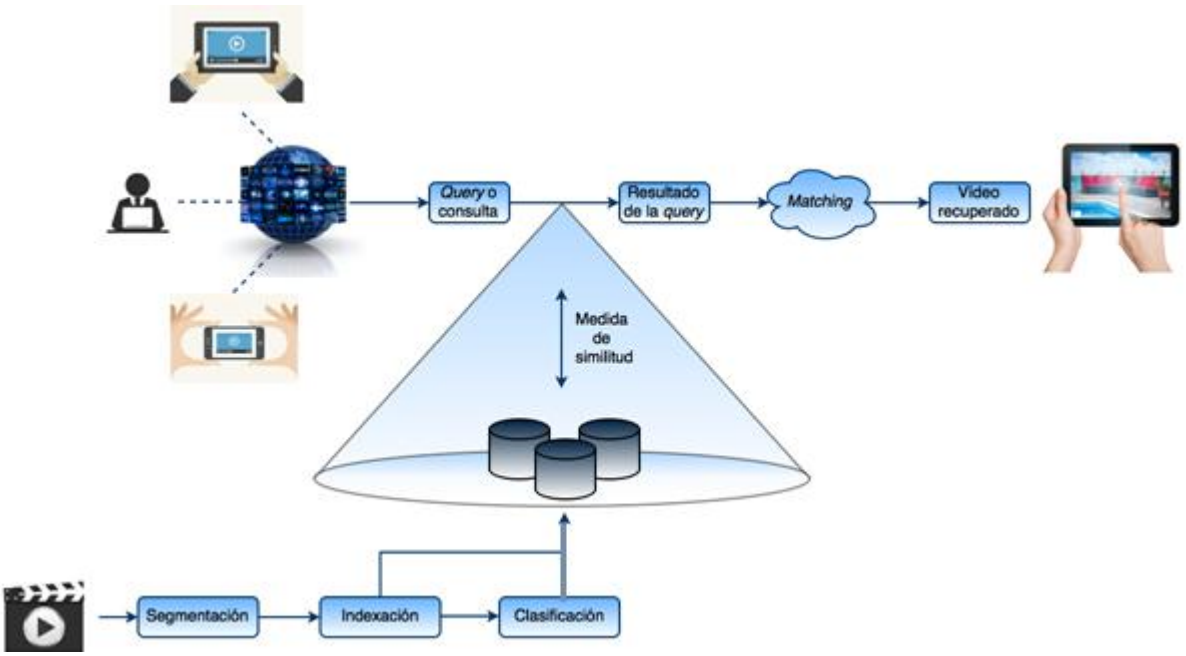

Figura 1. Procedimiento completo de indexación y recuperación mediante búsqueda de vídeo por etapas. Fuente: Propia.

En la parte superior, se representan a distintos usuarios, que a su vez acceden desde diversos dispositivos como ordenadores, tabletas o *smartphones*, que pretenden buscar algún vídeo

almacenado en internet. Para ello, realizan una consulta o *query* a través de algún navegador en un motor de búsqueda, del que se obtiene un conjunto de resultados. Dentro de este conjunto de vídeos obtenidos, que ya han sido seleccionados por un filtro previo (se devuelven varios vídeos cuyas características concuerdan con las proporcionadas en la *query*), el usuario escoge aquel que se asemeja más a su petición o que posee las características buscadas.
Por otro lado, en la parte inferior, para que un vídeo pueda ser encontrado por un usuario es necesario que previamente se hayan seguido las sucesivas etapas: segmentación de su contenido, indexación o anotación de sus características más representativas y clasificación en base a las mismas. Hecho esto, el vídeo es subido a la red y almacenado en grandes bases de datos. Tras este proceso, ya sería posible proceder con la comparación de similitud (*matching*) entre las características deseadas, que han sido introducidas por el usuario, y las que contiene el vídeo recuperado. Concretamente, se reconoce por *matching* al conjunto de vídeos que se recuperan como resultado de la *query* introducida, vídeos que crearán algo similar a un *ranking,* en función del grado de similitud que tengan con ella.

A continuación, se desarrollan las principales ideas contenidas en las diferentes etapas y que permiten concluir en la correcta indexación de material multimedia:

2.3 Segmentación

La segmentación es el primer paso a llevar a cabo cuando se persigue la identificación de unidades homogéneas en una secuencia de vídeo para su posterior indexación.
Aun así, el principal objetivo que se persigue cuando se busca indexar un vídeo es abstraerse de las particularidades de cada unidad homogénea y dar un sentido global al vídeo del caso de estudio, de modo que se obtenga una o varias etiquetas generales que lo identifiquen y diferencien del resto.

Los vídeos están estructurados de acuerdo a una jerarquía descendiente del vídeo completo, que se divide de forma consecuente en *scenes*, *shots* y *frames*. Durante la segmentación se realiza un análisis de dicha estructura, el cual consiste en la segmentación de las distintas escenas, la detección de las fronteras o los límites de cada *shot* y, por último, la segmentación en *frames* y la extracción de los *key frames*.

Gracias a esto se puede realizar una descomposición de vídeo en distintos niveles de granularidad [3][4][6], lo que facilita el análisis de las características, siguiente paso en el camino de la indexación dinámica. Se observan los citados niveles en la Figura 2.2 mostrada a continuación.

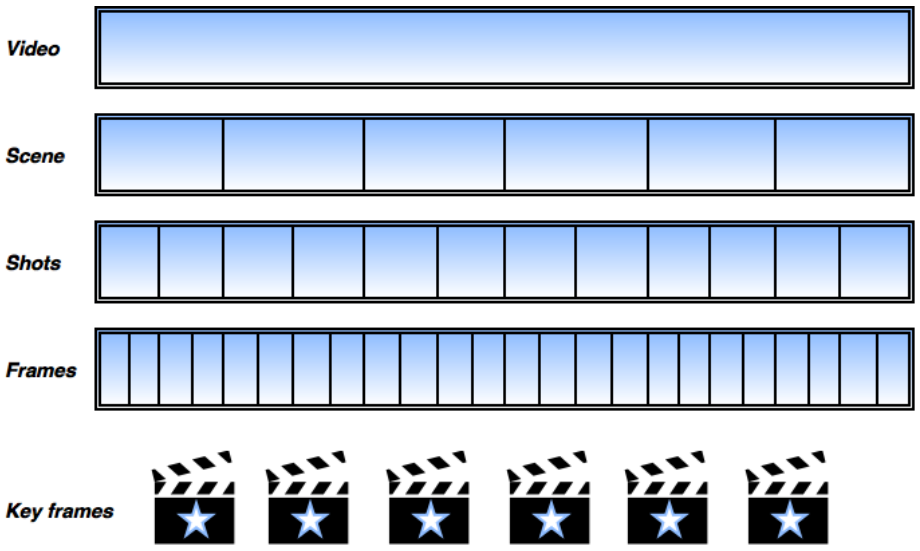

Figura 2. Representación jerárquica de un vídeo. Fuente: Propia

Definiendo los conceptos más importantes, se tiene que un ***shot*** se define como una secuencia de imágenes que presenta una acción continua, la cual es captada a partir de una única operación en una sola cámara. Los *shots* pueden considerarse como la unidad de indexación más pequeña donde no hay cambios que puedan ser percibidos en el contenido de la escena y los conceptos de más alto nivel a menudo se construyen mediante la combinación y el análisis de las relaciones inter e intra *shot*.
Por su parte, se conoce como ***scene*** a aquel conjunto de *shots* que claramente pertenecen al mismo tema. Además, las escenas tienen un nivel superior de contenido semántico. Para facilitar la segmentación por escena, ésta se basa en agrupar *shots* de acuerdo al contenido de imágenes, texto y audio.

Por último, los distintos ***frames*** son cada una de las imágenes en las que se descompone un vídeo. Para cada *shot*, se puede encontrar la denominada *key frame* o imagen clave, muy representativa del *shot* en cuestión por representar de manera global a su contenido y a partir de la cual se extraerá las características.

Una puntualización importante en este tema es que, si se pretende obtener características que incluyan información temporal (como el audio o el movimiento) la extracción a partir de la *key frame* no es una opción válida; por lo que la primera solución sería extraer este tipo de información a nivel de *shot* (como mínimo, pudiendo ascender en la escala jerárquica establecida).

Como puede observase, la calidad de la indexación del vídeo va a estar ligada, en gran medida, a la eficiente segmentación del mismo a nivel de *shot*, ya que es éste el que permite diferenciar a grandes rasgos las distintas unidades homogéneas en un vídeo. Por ello, un *shot*

es considerado como la unidad básica por medio de la cual analizar y organizar el contenido de un vídeo. En un determinado *shot* se pueden diferenciar los límites o fronteras, observando el comienzo y el final de la acción que engloba. Además, en los *frames* de un mismo *shot*, la relación que existe en el contenido semántico entre ellos es muy grande, por lo que es un objetivo claro el aprovechar esa redundancia para simplificar el análisis.

El estado del arte en cuanto a la **detección de *shots*** se puede agrupar en diferentes métodos, como los establecidos en [3][7], en los que se mencionan sus cualidades y principales desventajas:

1) **Aproximaciones basadas en la identificación de umbrales (parametrización).**

En esta técnica se establece un umbral en algún factor en el que se encuentra una similitud más o menos constante entre los distintos *frames* del *shot*. El límite o la frontera se detecta cuando el valor de ese factor decae más allá del límite establecido como umbral. Este umbral puede ser de varios tipos: global, adaptativo o una combinación de ambos.

Ventajas y desventajas: Esta aproximación resulta bastante efectiva en términos generales de su aplicación. Sin embargo, requiere una visualización, análisis y estudio previo del contenido a umbralizar, de cara a establecer unos valores razonables que puedan asegurar un buen funcionamiento del algoritmo. Aun así, es una aproximación útil cuando se conoce bien la información a indexar, ya que permite modificar parámetros y valores concretos de cara a ajustar el algoritmo para un mayor rendimiento.

2) **Aproximaciones basadas en el aprendizaje estadístico.**

En este enfoque, los *frames* se clasifican en función de la base de las características comunes, dependiendo de su variación. Para ello se hace uso de algoritmos de aprendizaje tanto supervisado, como lo son *SVM* o *Adaboost*, como no supervisado, por ejemplo *K-means*.

Ventajas y desventajas: Este otro enfoque suele ser muy efectivo, ya que no requiere intervención (salvo la fase de entrenamiento previa a cualquier algoritmo basado en el aprendizaje máquina) por ser el propio algoritmo el que aprende de manera autónoma tras dicho aprendizaje. Algoritmos o sistemas de este estilo pueden acarrear problemas como el sobreajuste, perdiendo el carácter general y llevando a resultados erróneos.

3) **Aproximaciones basadas en transiciones graduales.**

Esta otra aproximación pone especial atención a la detección de transiciones graduales, lo que lleva a un análisis en múltiples resoluciones. Esta técnica es reciente y tiene diferentes enfoques, como puede ser aquellas que se basan en las transiciones de acuerdo a la curva de distribución de la varianza de la información de los bordes (*edges*) en distintos *frames*.

Ventajas y desventajas: Se trata de una aproximación con una gran utilidad, ya que permite adaptar el algoritmo a distintas casuísticas, lo que aumenta la eficacia de un modo considerable con aquellos cuyo contenido es conocido a priori. Sin embargo, como en la mayoría de los casos, para vídeos cuyo contenido no es conocido de antemano, haría falta una etapa de entrenamiento. Aun así, dadas las variadas posibilidades que presenta este enfoque, con el entrenamiento adecuado del algoritmo podrían llegarse a conseguir resultados muy favorables.

2.4 Indexación

Hasta hace unas décadas, la construcción de índices de vídeo solía realizarse manualmente, asignando un número limitado de palabras claves en relación con el contenido del vídeo analizado. Esta tarea conllevaba unos costes altos y un gran consumo de tiempo, por lo que se vio la necesidad de automatizar esta tarea. Este mecanismo se conoce como indexación de vídeo (*video indexing)*, definido como el proceso de etiquetado automático del vídeo.

Entre las primeras decisiones que deben tomarse para hacer posible la indexación de un vídeo se encuentra si ésta se hará o no por **contenido semántico** [5], uno de los enfoques existentes más utilizados y que se encuentra modelado en tres niveles diferentes:

- **Propósito:** Es el nivel más alto (*top level*). Se basa en la observación de que un autor crea un vídeo con un propósito determinado y, por tanto, todos los vídeos que compartan el mismo propósito tendrán una **intención similar** en lo que quieren transmitir. Por ejemplo, los anuncios publicitarios están motivados para que un cliente compre un producto concreto; un vídeo sobre un noticiario persigue un fin informativo, con el fin de comunicar una noticia o información.

- **Apariencia o diseño:** El siguiente nivel se relaciona con aquellos vídeos que poseen un **estilo y contenido similar**. Partiendo de este nivel se pueden diferenciar géneros y subgéneros en el análisis de una colección de vídeos. Por ejemplo, aquellas noticias que siempre muestran de fondo un mapa físico y su predicción del tiempo se clasifican como noticias de predicción meteorológica; o aquellas que muestran tablas de resultados deportivos, se identifican como noticias de deporte.

- **Partes del contenido:** Constituyen aquellos eventos que tienen un **significado completo por sí mismos** al unirse, lo que se conoce como **unidades lógicas**, y que no cambian durante un período de tiempo (normalmente breve). Dentro de los eventos se encuentran ejemplos tales como una explosión en una película de acción, un gol en un partido de fútbol o las cotizaciones de bolsa en una emisión de noticias financieras

Tras esta aproximación, se remarca la variedad de elementos de estilo tanto visuales, de audio y textuales que pueden dar forma a los distintos niveles de indexación por contenido semántico y que contribuyen a expresar la intención del autor.

Los colores, la iluminación, el encuadre de la cámara, distancia o movimiento general de la imagen son parte de los elementos de estilo presentes en la modalidad visual, mientras que en la de audio puede diferenciarse el ritmo, diferentes propiedades en la música, volumen o intensidad. Por su parte, en el contenido de texto se encuentran las expresiones o el estilo de escritura, entre muchos otros.

Si bien es cierto que los métodos más comunes siguen este tipo de estructura para comenzar con la indexación de los vídeos bajo análisis, las posibilidades son infinitas y pueden adaptarse en gran medida al fin que se persiga, ya sea una clasificación más genérica o enfocada de forma particular a diferenciar algún tipo de evento en concreto. Por esta razón, existirá mucha variedad en los posibles diseños e implementaciones de indexación.

Antes de continuar, se pretende definir varios términos que pueden ser ambiguos:

- **Característica:** Rasgo que se extrae, que sirve para definir y diferenciar.
- **Índice/Etiqueta/Anotación:** Valor que se atribuye, utilizados como sinónimos.

Además de lo comentado anteriormente, en el momento de indexar surgen tres cuestiones importantes [5]:

1) ¿Qué indexar?

Se relaciona con la **granularidad**. Por ejemplo, si se realiza la indexación de un vídeo al completo, a nivel de *shot* o para cada una de las *frames* de forma individual.

2) ¿Cómo indexar?

Se relaciona con las distintas modalidades y el gran rango de **técnicas y algoritmos** que pueden utilizarse como método de indexación.

3) ¿Qué tipo de descriptor utilizo?

El índice que se generará (a partir de las características que se extraigan) hace referencia al **tipo de información que pretende registrarse**. Por ejemplo, en un vídeo que contenga un partido de fútbol, podrían etiquetarse los nombres de los jugadores que aparecen en la imagen, la posición dentro del campo de juego en un momento determinado, o incluso por ambas.

Una vez decididas estas primeras cuestiones, se procederá a la extracción de las características adecuadas para ser capaces de obtener una información relevante según la cual indexar, de acuerdo al objetivo que se persigue.

2.5 Extracción de características

La cantidad de características extraíbles en un vídeo es realmente elevada y queda determinada por el nivel de detalle que se pretenda obtener (*accuracy*), por lo que la dificultad que tendrá el algoritmo residirá en la medida en la que se pretenda llegar a la consecución de este objetivo.

Es por ello que en esta etapa debe seleccionarse qué características serán de mayor utilidad, en numerosas ocasiones ligada al conocimiento del tipo de vídeo bajo análisis, y con qué algoritmo, técnica o método podrán extraerse de una manera más eficaz, de modo que la calidad del algoritmo resultante sea la mayor posible.

De forma general, las características a extraer pueden diferenciarse en dos grandes bloques: características espaciales y características temporales. Además, es común agregar un bloque más que aúne características que tienen cabida en ambos bloques de la distinción (espacio-temporales):

<table>
<tr><td rowspan="3">Características</td><td rowspan="2">Espaciales</td><td>Globales u holísticas</td><td rowspan="2">→ Color.
→ Textura.
→ Formas y bordes.
→ ...</td></tr>
<tr><td>Locales o por partes</td></tr>
<tr><td colspan="2">Temporales</td><td>→ Información de audio.
→ Información de movimiento.
→ ...</td></tr>
<tr><td colspan="2">Espacio-temporales</td><td>→ Movimiento estudiado por zonas.
→ Volúmenes.</td></tr>
</table>

Tabla 1. Clasificación de tipo de características. Fuente: Propia

Mediante estas propiedades se puede realizar un estudio de la imagen a bajo nivel con sus características más significativas. Para aplicar cada una de ellas, existen en el estado del arte diferentes técnicas como histogramas de color globales y locales (GCH, LCH, respectivamente), filtros de textura *Gabor*, aproximaciones mediante el gradiente, histogramas de movimiento,

flujos ópticos, etc. Si además, se incluyen las características asociadas a los objetos presentes en el vídeo, el procedimiento de extracción de características resulta más efectivo.

Se profundiza en estas y otras técnicas específicas de extracción de características para vídeos de noticiarios en el Capítulo 3.

Obtenidas las características, se procede a realizar un *matching* o, dicho de otro modo, un emparejamiento de los tipos de índices que se han decidido fijar con las características extraídas. Este proceso se realiza de manera cualitativa, tras barajar las posibles combinaciones de características que pueden resultar en determinados índices; es decir, se definen unas reglas semánticas que asocien características con índices para la clasificación.

Estos índices pueden ser vistos como **etiquetas**, de modo que a un vídeo concreto quede sesgado por distintas características resumidas en palabras clave que, además, serán intencionadamente atribuidas en función de lo que un usuario puede querer obtener en su búsqueda de un vídeo concreto.

Al igual que la mayoría de etapas, en la indexación de vídeo existen distintos niveles de profundización en el algoritmo. En el caso de los índices o etiquetas se pueden establecer muchos niveles de detalle, desde tan básicos como son interior/exterior, noche/día; aumentando el nivel de detalle como sería distinguir entre naturaleza/urbano, a otros mucho más rigurosos, que diferencian clases particulares dentro de un evento concreto, como distinción entre deportes fútbol/baloncesto. Incluso, podría crearse a un más alto nivel distinciones más concretas como fútbol/rugby o natación/waterpolo, en los que se debe ahondar con bastante más detalle en sus diferencias para ser capaz de obtenerlos adecuadamente.

2.6 Clasificación

La clasificación de vídeos ayuda a incrementar la eficiencia en su recuperación y es una de las tareas más importantes que se debe valorar a la hora de trabajar con material multimedia almacenado en grandes repositorios.

Previamente al proceso de clasificación, la información es extraída de los componentes presentes en el vídeo para tras ello poder situarlos en diferentes categorías que se definen en base a ciertos criterios.

Dado que esta etapa puede venir definida por un gran número de aproximaciones diferentes, no se profundiza en las posibles clasificaciones (en el capítulo 3 se comentará la idea de

clasificación generada para nuestro propio algoritmo), las cuales quedarán definidas y modeladas según los criterios subjetivos del usuario que indexa con un objetivo final.

2.7 Recuperación

Para la búsqueda, filtrado, selección, exploración y manipulación de los documentos de vídeo, se requiere la descripción de los mismos mediante un índice basado en su contenido, como ya se ha visto que se realiza en etapas previas. Esta técnica se conoce como *Content Based Video Retrieval* (CBVR) [2].

Es importante tener en cuenta que es al finalizar la etapa de la indexación cuando ya los vídeos pueden ser subidos a los repositorios, dado que tras esta fase ya se encuentran etiquetados con anotaciones sobre sus características, lo que permitirá su posible recuperación mediante los filtros adecuados.

Una vez han sido obtenidos los índices de los vídeos, la recuperación de los mismos puede ser llevada a cabo. Tras recibir la *query*, el algoritmo de medida de similitud (*matching*) entra en juego, buscando aquellos candidatos que se corresponden con la *query* ejecutada.

Existen distintos tipos *query* en función de cómo se han indexado los vídeos [1][2][7], entre los que se diferencian dos grupos de forma general, como puede apreciarse en la Figura 2.3.

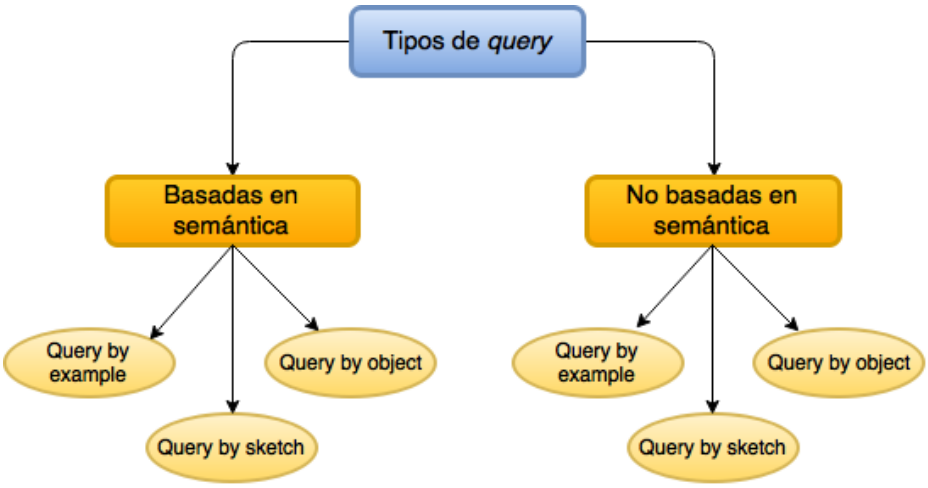

Figura 3. Tipos de *query*. Fuente: Propia

De entre los tipos pertenecientes a las basadas en **contenido no semántico** se tienen:

- ***Query by example:*** Utiliza en gran medida características de bajo nivel, sobre todo estáticas, obtenidas a partir de la *key frame* tomada como modelo.
- ***Query by sketch:*** Este tipo de *query* permite a los usuarios dibujar o seleccionar *sketches* (bocetos, esquemas) para representar aquellos vídeos que tratan de buscar. Así, las características extraídas de los *sketches* son comparadas con las características almacenadas en los vídeos de los repositorios.
- ***Query by object:*** Se proporciona en la consulta el objeto concreto del que se pretende encontrar similitud. Se devolverá, por tanto, las ocurrencias que coincidan junto con las posiciones o localizaciones de estos objetos en el video. Dentro de este tipo, se pueden diferenciar otros más concretos como aquella que asume caras por objetos (*query by faces*).

Mientras que, las basadas en **contenido semántico** se caracterizan por:

- ***Query by text* o *query by keywords:*** Uno de los más populares en la recuperación de vídeo basada en contenido, por tratarse de un método simple y eficaz como aproximación para identificar vídeos en repositorios. La *query* se representa mediante un conjunto de palabras clave o *keywords*, que pueden hacer referencia a metadatos, contenidos visuales, etc.
- ***Query by shot:*** Algunos sistemas utilizan el *shot* completo dentro de un vídeo como *query* en lugar de su *key frame*. Los resultados obtenidos pueden ser muy buenos pero en general no compensa, debido al elevado coste computacional del que hace uso.
- **Combinación de distintos tipos de *query:*** Como método opcional, puede utilizarse la combinación de varios tipos de los descritos anteriormente para un mayor grado de precisión en la recuperación. Por tanto, se trata de un tipo de *query* multimodal.

Los resultados devueltos pueden ser optimizados mediante un *feedback* adecuado (conocido como ***relevance feedback***), donde los vídeos obtenidos en respuesta a la consulta se sitúan en un ranking que se crea por el usuario en tres categorías diferentes según el grado de participación: explícito, implícito y *pseudofeedback*.

Este ranking se genera de cara a someter al algoritmo a un refinamiento que, alimentado con este *feedback*, sea capaz de precisar mejor en la próxima búsqueda. Entre estos métodos de refinamiento se encuentra la optimización de consultas por punto (*query point optimization*), el ajuste del peso de las características (*feature weight adjustment*) y la incrustación de información (*information embedding*).

2.8 Conclusiones

A modo de conclusión, en este capítulo se han recogido y se han descrito las distintas etapas del ciclo de vida de un sistema de indexación automática y recuperación por consulta de vídeo, haciendo hincapié en las técnicas más comunes dada su eficacia.

Es importante destacar que, como en muchos otros ámbitos, el diseño de estos tipos de sistemas puede ser muy variable, incorporando otras funcionalidades y/o características más enfocadas al objetivo concreto que se persiga.

Capítulo 3. Indexación de noticias

3.1 Introducción

En este capítulo se realizará un estado del arte más enfocado al tema principal de este trabajo, el cual gira en torno a los programas informativos, también conocidos como noticiarios [8], que presentan un amplio campo de estudio por su capacidad de ser adaptados a distintas definiciones de modelos de segmentación, extracción e indexación.

Así mismo, el tipo de vídeos con el que se trata presentan un contenido muy variado que engloba un gran abanico de contextos y temáticas. Dada esta casuística, es posible realizar un sistema de indexación y recuperación a distintos niveles de granularidad, según se pretenda ahondar más o menos en el nivel de detalle.

3.2 Segmentación en unidades homogéneas

Recordando lo explicado en el capítulo 2, la jerarquía de descomposición de un vídeo en unidades más simples sigue una estructura regular y descendente en escenas, *shots* y *frames*; y los modelos diseñados para realizar esta segmentación varían en gran medida por el tipo de vídeo que se quiera analizar.

La segmentación en las diferentes unidades que se pretendan diferenciar en el vídeo es una de las tareas más complicadas de llevar a cabo. Esto no es necesariamente por la complejidad técnica a la hora de realizar las divisiones entre distintas unidades, sino por el gran número de criterios que pueden tenerse en cuenta a la hora de segmentar el vídeo en unidades homogéneas y la cantidad de aproximaciones que pueden desarrollarse para conseguirlo.

Los principios utilizados para definir las escenas o los *shots* varían extensamente en función de si tratan con programas de noticiarios, documentales, películas, etc. La estructura rígida que presentan los noticiarios los hace idóneos para definir una aproximación y procedimiento específico para la indexación y recuperación de sus vídeos [8], ya que hace que sea abordable la detección de objetos en la escena y la segmentación por el contenido semántico de la misma. El conocimiento de esta estructura permite, del mismo modo, el análisis de parámetros característicos y su posterior extracción persiguiendo el resultado más óptimo. Por esta razón, el estudio principal de este trabajo girará en torno a este género, siguiendo la línea de muchos investigadores recientes que encuentran en este campo una motivación adicional.

Mediante la observación del contenido disponible dentro de un noticiario se puede llevar a cabo la clasificación, a nivel escena o a nivel *shot* en distintas categorías. Dentro de estas categorías podrían diferenciarse tres grandes bloques:

1. Presentador y otros planos de **estudio.**

En este bloque se incluirían todas aquellas unidades o segmentos de vídeo que contienen encabezamientos de los titulares, acompañados de una introducción a la noticia por parte del presentador.
También es posible encontrar dentro de esta categoría otros planos de estudio que incluyan entrevistas o debates entre distintos individuos.

2. **Reportajes** en directo o pregrabados (remotos).

Muestran el contenido de la noticia, los hechos que se relatan en diferentes localizaciones. Este reportaje, normalmente narrado por un periodista que aparece en ocasiones en la imagen, acumula un gran contenido del que extraer características.

3. **Logos/animación.**

Incluye información tanto visual como de audio (por ejemplo la sintonía del informativo).
Del mismo modo también se incluyen en esa categoría anuncios y *sketches* publicitarios.

Hasta ahora, la mayoría de los algoritmos de segmentación de vídeo se basaban simplemente en explotar la información visual. Sin embargo, se ha observado que la información visual por sí sola obtiene resultados menos precisos que cuando se combina con información más amplia[7][8][9].

El uso de múltiples ángulos de cámara o efectos especiales complica aún más esta tarea. Como resultado, los cambios de *shot* están mal clasificados como cambios de escena en muchos casos. Típicamente, un método eficiente para separar un cambio de escena de un cambio de *shot* es mediante el empleo de la información de audio. Un cambio de escena está más a menudo acompañado de un cambio significativo en las características de audio, mientras que un cambio de shot no muestra una diferencia sonora tan distintiva. Un ejemplo notable es un anuncio de televisión típico, que constituye una sola escena con características similares de audio, pero que a su vez está compuesto por numerosos *shots*. En este punto, la definición de una unidad semántica que tome el audio o el sonido en la escena como complemento indispensable es esencial. Por ello, se define a un segmento de audio como unidad semánticamente coherente que se distingue por las características básicas del sonido dominante. Este segmento puede ser detectado cuando la mayoría de las fuentes dominantes cambia.

Por último, quedaría mencionar la información relativa al texto y su utilidad en esta etapa de segmentación. Si bien es cierto que, explotada de la manera adecuada y estableciendo reglas y pautas tras el previo análisis, puede derivar en un aumento del rendimiento y eficiencia del sistema, también lo es que este tipo de información textual es sobre todo más ventajosa a la hora de identificar características en ella, por lo que el reconocimiento del texto está más enfocado a la etapa de extracción.

A continuación, se muestra en la Figura 4 un resumen de las distintas categorías que pueden diferenciarse en el proceso de abstracción para la segmentación. Para ello, se utilizan los vídeos del conjunto de estudio en el que se realizarán las pruebas:

Figura 4. Categorías de clasificación en noticiarios. Fuente: Propia

3.3 Indexación

Dado que los noticiarios son eventos televisivos que ocurren diariamente, normalmente incluso más de una vez al día, las bases de datos que almacenan su contenido multimedia no pueden permitirse realizar una tarea de indexado de forma manual. Igualmente, un usuario que pretende recopilar archivos de vídeos sobre un tema en concreto, no puede afrontar la tarea de buscar el material deseado en un repositorio tan extenso sin disponer de un método automático que le facilite el trabajo.

Para construir un sistema que sea capaz de lograr este objetivo, la etapa de extracción de características (*features*) es clave, ya que en función de su precisión y la rigurosidad con que sean extraídas influirá de manera directa en los resultados obtenidos.

3.3.1 Extracción de características en función del tipo de contenido

El principal propósito de esta etapa es extraer características distintivas en la unidad homogénea por la que se haya decidido segmentar al vídeo bajo estudio.

Normalmente, esto vendrá dado a partir de la *key frame* si se trata de características estáticas o, si por el contrario, se precisa información temporal (movimiento, audio) esta información se obtendrá como mínimo, a nivel de *shot*, siendo también bastante común a nivel de escena [9].

Se procede a hacer un recopilatorio de las principales características visuales que se suelen extraer de modo global para un indexado genérico de vídeo. Se hace hincapié en la generalidad de las siguientes características, ya que constituyen parámetros a bajo nivel que con su combinación pueden dar fruto a muchas aproximaciones diferentes:

3.3.1.1 Características espaciales

Color

El color es una característica muy importante en cualquier imagen o vídeo. A partir de ella, pueden ser extraídas numerosas características y atributos inherentes a la imagen o vídeo que se está analizando. Por esta razón, el análisis del color en la representación de una imagen va a cobrar una importancia vital, al igual que su tratamiento será uno de los puntos más fuertes a tratar. De esta manera, se podrá hacer uso del color para representar a la imagen por sí misma o, combinándola con otras características, ser capaz de generar una definición propia que la caracterice de manera particular, permitiendo su indexación y posterior recuperación.

La importancia del color en una imagen para este fin que se propone se debe principalmente a los siguientes factores [9][10]:

1. El **ojo humano es muy sensible al cambio de color** y es una de las características fundamentales que permiten reconocer imágenes.
2. El **color es invariante** respecto a la escala, translación y rotación de la imagen.
3. El color es una característica **muy representativa del contenido** de la imagen.

Se plantean a continuación alguna de las técnicas de extracción de color más generales. Al mismo tiempo, se trata de **técnicas relativamente fáciles de aplicar** y las cuales ofrecen **resultados buenos y adaptables** a muchos ámbitos diferentes.

Técnicas en el color:

Concretamente, se puede realizar una primera aproximación a las técnicas del color haciendo una distinción en tres pilares básicos, como la que se propone en [11]: descriptores de color, histogramas de color y correlogramas de color. Estas clasificaciones están altamente relacionadas y pueden utilizarse conjuntamente en la extracción de las propiedades del color, como se describe a continuación.

Descriptores de color

Los descriptores de color en imágenes y vídeos pueden ser **globales o locales**. Los descriptores globales especifican el contenido de color en todo el conjunto de una imagen, pero no aportan información sobre su distribución espacial. En cambio, los descriptores locales hacen referencia a regiones particulares de la imagen, lo que permite describir su disposición espacial. Por ejemplo, los descriptores de color MPEG-7 consisten en un número de descriptores de histograma, un descriptor del color dominante y un descriptor de diseño de color (*Color Layout Descriptor* o CLD).

Histogramas de color

Un histograma de color:

$$h(imagen) = h_k(imagen), k = 1, \dots, K$$

es un vector K-dimensional tal que cada componente $h_k(imagen)$ representa el **número de píxeles relativos a un color** $\boldsymbol{Ck}$ en la imagen o, lo que es lo mismo, la fracción de píxeles que son más similares al color $\boldsymbol{Ck}$ correspondiente.

Para construir el histograma de color, la imagen de color debe ser transformada a un espacio de color apropiado (el más común, RGB) y cuantificada de acuerdo al diccionario particular de tamaño K.

Correlogramas de color

Un correlograma de color de una imagen es una tabla indexada por pares de colores, donde la entrada de orden $k(i,j)$ especifica la probabilidad de encontrar un píxel de color *x* a una distancia de un píxel de color *y* en la imagen. Es decir, un correlograma constituye una **imagen de correlación de estadísticas**.

Este método resulta robusto para tolerar grandes cambios en la escena, como pueden ser cambios en las posiciones de visualización, cambios de fondo, oclusiones parciales, *zoom* de la cámara, etc. Esta característica destila la correlación espacial de colores y es a la vez eficaz y de bajo costo para la recuperación basada en el contenido de la imagen.

Además, a diferencia de propiedades puramente locales como la posición del píxel o la dirección del gradiente, y de propiedades puramente globales, tales como la distribución del color, los correlogramas tienen en cuenta la correlación espacial del color local, así como la distribución global de esta correlación espacial.

Es importante destacar que el color está generalmente definido en un espacio tridimensional y existen distintos espacios en los que puede ser tratado.

El espacio de color por antonomasia es el RGB, definido como un cubo con ejes rojo, verde y azul, por lo que un color en RGB se representa como un vector de tres coordenadas. Cuando todos los valores están fijados a 0, el color correspondiente es el negro. Por el contrario, si todos los valores se fijan a 1, el color correspondiente es el blanco.

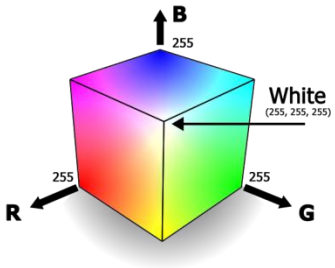

Figura 5. Representación del espacio de color más común, RGB. Fuente: Propia

Los histogramas de color se definen como un conjunto de barras donde cada barra denota la probabilidad de los píxeles de la imagen para cada uno de los tres ejes. Teniendo en cuenta el descriptor utilizado, se diferencian dos tipos [9][10][11][12]:

- Descriptor de histograma de color global, GCH (Global Color Histogram).
- Descriptor de histograma de color local, LCH (Local Color Histogram).

Además, existen otros enfoques más recientes como aquel que parte de GCH pero es acumulativo, conocido como histograma acumulado de color global (CGCH) y el descriptor de coherencia de vectores de color, CCV (Color Coherence Vector).

Todos estos son descriptores populares de la literatura y que por lo general se utilizan como línea de base para las comparaciones, siendo simples y efectivos.
La principal diferencia entre LCH y GCH es que el primero sí incluye información acerca de la distribución de color por regiones de la imagen, lo que lo hace más preciso. Para ello, el primer paso es dividir la imagen en bloques y obtener un histograma de color para cada uno de ellos, calculando luego la diferencia entre imágenes según estos histogramas bloque a bloque. La distancia total será determinada como la suma de todas esas distancias obtenidas mediante algún algoritmo, típicamente la raíz cuadrada de la distancia euclídea.

El método LCH soluciona algunas de las principales desventajas del GCH al comparar por regiones de la imagen. Sin embargo, debido a que compara regiones en la misma localización, el método no funciona bien cuando se producen rotaciones o traslaciones en dicha imagen.

Por último, los descriptores basados en correlogramas son prometedores porque codifican información espacial, como ACC (*Auto Color Correlogram*) o JAC (*Join Auto Correlogram*), la cual suele perderse al emplearse histogramas de color más simples.
Dada la gran importancia del color en la extracción de características en una imagen de vídeo existen numerosas implementaciones que analizan este atributo de diversas formas. Por ello, existen muchas otras técnicas que analizan el color en diferentes perspectivas. Algunas de ellas se basan en el momento de color o en modelos Gaussianos.

Aplicaciones en el color:

La información de color puede ser muy distintiva para ciertos contextos en los vídeos de noticiarios, cuyos *datasets* utilizamos como caso de estudio.

En primer lugar, la información de color que se obtiene en un escenario de interior es muy diferente a uno de exterior, donde la **luminancia es un factor clave**. Además, es fácil distinguir algunos contextos dentro del ámbito de un noticiario por la información de color que puede detectarse en él.

Por ejemplo, para la sección de deportes, es lógico pensar que cuando el histograma concentre gran parte de su distribución de probabilidad en el eje de coordenadas

correspondiente al verde, pueda tratarse de un reportaje de fútbol, o rugby, cuyo campo se presenta de este color; si esto ocurre con el azul, podrá ser un deporte acuático, mientras que un color que concentre más probabilidad en el rojo podría tratarse de un deporte en pista interior como el baloncesto.

Textura

La información sobre la textura en una imagen es una propiedad muy característica de su contenido. Existen en la literatura varias propuestas en cuanto a su definición, siendo algunas de ellas las siguientes [10][12]:

1. "Su estructura se atribuye simplemente a los patrones repetitivos en la que los elementos están dispuestos de acuerdo con un regla de colocación".
2. "La textura en una imagen es descrita por el número y tipo de sus elementos y la organización espacial o la disposición de los mismos".

En una definición que toma a la textura de una manera general, esta puede ser entendida como un conjunto de variaciones de intensidad que siguen ciertos patrones repetitivos.

A diferencia del color, la textura es difícil de ser analizada teniendo en cuenta el valor de un sólo píxel, ya que se produce principalmente por la variación en un vecindario de píxeles. Esto hace que sea posible nombrar algunos atributos relativos a la textura como el contraste, la uniformidad o regularidad, la aspereza y la densidad en una imagen. Estos atributos son fácilmente observables y reconocibles en una imagen como arena, nubes, ladrillos, hojas, etc.

Es importante destacar que la mayoría de descriptores de textura trabajan en la escala de grises, existiendo sólo algunos que describen cómo se debería adaptar e incluir la información de color en la textura para mejorar el rendimiento del descriptor.

Técnicas en la textura:

Existen varias técnicas por las que analizar esta propiedad, pudiendo enumerarse cuatro grandes bloques [10]: métodos estadísticos, métodos geométricos, métodos basados en modelos y métodos basados en procesamiento de señales.

Métodos estadísticos

Una de las formas más tradicionales para analizar la distribución espacial de los niveles de gris en una imagen es por análisis estadístico, por ejemplo, mediante el cálculo de la probabilidad de co-ocurrencia de valores de gris en diferentes distancias y orientaciones. Las estadísticas se pueden calcular sobre los valores de los píxeles individuales (estadísticas de primer orden) o

sobre el valor de pares de pixeles (estadísticas de segundo orden). Aquellos que caracterizan texturas por medio de histogramas también utilizan métodos estadísticos. Uno de los métodos estadísticos más populares es la matriz de co-ocurrencia COM (*co-ocurrence matrix*), donde más recientemente se ha añadido al descriptor la información de color, resultando en CCOM (*color co-ocurrence matrix*) .

Métodos geométricos

Los métodos geométricos analizan texturas por "elementos de textura", también conocidos como primitivos. Este análisis se realiza teniendo en cuenta las propiedades geométricas de los primitivos, como el tamaño, la forma, el área y longitud. Habiendo siendo identificadas los primitivos en una imagen, se extraen de ellos reglas de colocación, como rejillas o vectores que unen los centroides de diferentes primitivos.

Este tipo de análisis se vuelve difícil para las texturas naturales, debido a que los primitivos y las reglas de colocación pueden ser irregulares (por ejemplo, el caso de las nubes en el cielo). Sin embargo, en la descripción de una pared de ladrillos esta técnica puede ser simple y con muy buenos resultados. El primer caso tomaría el elemento *nube* como primitivo, mientras que en el segundo caso, el elemento primitivo sería *ladrillo*.

Métodos basados en modelos

Los métodos basados en modelos se centran en la construcción de modelos de imagen que se pueden utilizar para describir y sintetizar texturas. Los parámetros del modelo capturan las cualidades esenciales que pueden percibirse y en las que el ojo humano puede diferenciar una textura clara. Por ejemplo, un punto más o menos brillante, texturas regulares horizontales o verticales, etc. Un ejemplo de método que sigue este modelo es el descriptor local de patrón binario.

Métodos basados en procesamiento de señales

Estos métodos de procesamiento de señales caracterizan texturas mediante la aplicación de filtros espaciales y frecuenciales sobre la imagen. Descriptores que siguen este enfoque son los basados en *wavelets* y filtros de *Gabor*.

Aplicaciones en la textura:

La extracción de características por textura de la imagen resulta muy eficiente para imágenes, sobre todo, de naturaleza, ya que en ella hay presentes muchos elementos reconocibles mediante su análisis por textura. Por esta razón, puede utilizarse de forma particular en nuestro estudio para reportajes de exteriores en noticias.

Un caso concreto donde podría tener aplicación sería en aquellos reportajes que informan sobre catástrofes provocadas por fenómenos naturales, en los que aparecen imágenes de agua, fuego y otros elementos reconocibles de forma relativamente sencilla mediante el análisis de su textura.

Shape and edge (forma y bordes)

Para poder obtener información sobre la forma y los bordes es necesario realizar una segmentación de objetos en la imagen. Esta información puede proporcionar una gran cantidad de datos útiles que pueden ser transformados en anotaciones relevantes en el momento de la indexación.

La forma de un objeto se refiere a su estructura física y puede ser representada por la región que ocupa, las fronteras o bordes que la definen, etc. Estas representaciones tienen una gran importancia en el reconocimiento de objetos. Esto se debe a que, a partir de una forma que se tiene como plantilla (lo que, tras revisar el estado del arte podríamos definir con la *query by object* específica), puede hacerse coincidir con otras formas, produciéndose un *matching* y devolviendo la similitud encontrada.

Existen dos tipos de aproximaciones para la detección de bordes y forma [9][11]: las que se basan en el contorno de la forma de los objetos y las que, además, tienen en cuenta el contenido interno a ese contorno, que suelen ser más complejas pero proporcionan mejores resultados.

Un punto importante a tener en cuenta es que los objetos que aparecen en la imagen son muy sensibles a cambios de rotación, traslación, escalados, etc. y que, por tanto, las técnicas aplicadas para extraer la característica relativa a su forma y bordes debe tener presente estos **factores de variancia**.

Además, un descriptor de forma debe ser capaz de realizar la recuperación de imágenes para el máximo tipo de objetos, no sólo para ciertos tipos de formas, por lo que debe ser **independiente de la aplicación**.

Técnicas en formas y bordes:

A continuación se detallan las técnicas utilizadas para la extracción de las características de borde y forma de los objetos. Antes de entrar en detalle con las clasificaciones de métodos existentes para este fin, se nombran algunos de los descriptores más simples que se utilizan para extraer esta característica:

- Área.
- Perímetro.
- Compacidad o circularidad
- Excentricidad.
- Alargamiento.
- Ortogonalidad.
- Orientación.

Yendo un paso más allá, se hace una clasificación general en función de si las características se extraen sólo del contorno o si por el contrario se extraen de la región completa (forma más contenido):

A su vez, cada método se divide en dos enfoques, el enfoque global (se basa en la forma representada como un todo) y enfoque estructural (si se representa por segmentos). La principal diferencia entre ambos es que el primero utiliza todos los píxeles contenidos en la forma del objeto para aplicar la técnica mientras que, en el segundo, se divide el objeto en subpartes.

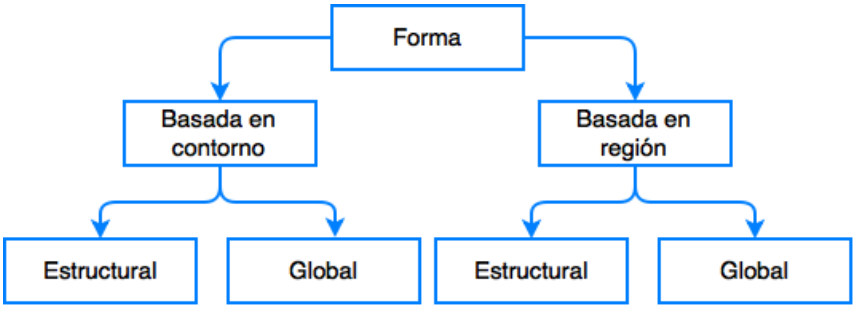

Figura 6. Clasificación de forma en sus distintas técnicas posibles. Fuente: Propia basada en [10]

Método basado en el contorno

Los métodos basados en el contorno son más populares que los métodos basados en la región, ya que está demostrado que los seres humanos son capaces de discriminar objetos principalmente por sus características de contorno. En algunas aplicaciones, el contenido interior no es importante, por lo que esta técnica es ampliamente utilizada para este tipo de aplicaciones.

Sin embargo, existen varias limitaciones inherentes a este método, ya que es sensible al ruido y a variaciones en la imagen (como rotaciones, traslaciones, escalados, etc.) por utilizar sólo una pequeña parte de la información, la información de contorno.

Uno de los operadores basados en el contorno y perteneciente al grupo reconocido como detectores de esquinas es el **operador Harris**, el cual lleva décadas siendo refinado desde que su autor lo definió en [13]. En realidad, no se centra sólo en localizar esquinas, sino más bien en cualquier ubicación de la imagen que tiene gradientes pronunciados en todas direcciones a una escala predeterminada. Por tanto, este detector es muy sensible a los cambios de escala de la imagen, por lo que no proporciona una buena base para la búsqueda de imágenes de diferentes tamaños.

Otro operador basado en el contorno es el conocido como **operador Canny**, utilizado principalmente en los histogramas de direcciones de bordes o HED (*Histogram of edges directions*). Este operador también se definió tiempo atrás en [14] y a día de hoy sigue utilizándose como principal herramienta de extracción de bordes.

Método basado en la región

Las limitaciones que presenta la anterior técnica pueden ser superadas por los métodos basados en la región. Estos métodos son más robustos, ya que utilizan toda la información disponible relativa a la forma y su contenido, por lo que se consigue una recuperación más precisa al trabajar adecuadamente en estas condiciones.

La mayoría de los descriptores basados en la región están enfocados a utilizar una metodología enfocada en los denominados **momentos**, los cuales hacen uso de propiedades estadísticas y reflejan por completo la orientación de un objeto tras calcular para él distintas funciones que lo describen y lo definen de forma invariante a propiedades como la rotación.

Aplicaciones en formas y bordes:

La información relativa a la forma y los bordes está muy ligada a la texturización que existe en la imagen, por lo que está característica se encuentra muy relacionada con la anteriormente comentada.

La textura se refiere al patrón estructural que se observa en la superficie de los objetos de la imagen (madera, arena, hierba, telas, etc.). Esta repetición de elementos básicos o *textels* pueden ser periódicos o deterministas (suelen corresponderse con elementos artificiales) o aleatorios (suelen corresponderse con elementos naturales), pero siempre mantienen el mismo

patrón de repetición en un amplio rango de la imagen bajo estudio. Entre los principales problemas que surgen al detectar la textura se encuentran el ruido y la oclusión, por lo que se produce en la forma una distorsión arbitraria que a menudo corrompe el reconocimiento de objetos.

En el ámbito de los noticiarios, el cual se toma en este trabajo como principal caso de estudio, la detección de ciertos objetos puede ser clave para diferenciar distintos escenarios. Por ejemplo, la identificación de un micrófono puede atribuirse a una entrevista.

Si bien es cierto, el número de objetos diferentes que pueden aparecer en un noticiario es muy elevado, sobre todo cuando se trata de reportajes sobre distintas cuestiones que no se relacionan por objetos en concreto. Por ejemplo, los objetos identificables en un reportaje sobre una catástrofe natural no tienen apenas relación con aquellos que podrían diferenciarse en un reportaje sobre un juicio por asuntos de fraude fiscal, lo cual puede establecer un criterio de clasificación a la hora de diseñar un sistema de indexado.

3.3.1.2 Características temporales

Las características que se basan en la información temporal que puede extraerse de ellas tienen tres principales ventajas [15]:

1. La información temporal está bien definida.
2. La información temporal puede ser normalizada.
3. La información temporal permite organizar jerárquicamente aquello que se representa.

Esta información temporal es dinámica y puede ser analizada de muchas formas diferentes.

A continuación se centra el estudio de las características temporales de acuerdo a dos pilares: el movimiento presente en la imagen y las características de sonido o audio que pueden extraerse en ella, las cuales ofrecen una gran y amplia información.

Información de movimiento

El movimiento es una propiedad intrínseca del mundo y una parte integral en la experiencia visual. Cuando se trabaja con contenidos multimedia, el movimiento constituye una rica y amplia fuente de información en la organización perceptiva, el reconocimiento de objetos y la comprensión de una escena [15].

La extracción de características de movimiento se basa en el seguimiento de las entidades de interés presentes en la imagen y el análisis de su movimiento aparente. Esta técnica de seguimiento deberá hacer frente a una variedad de situaciones que presentan desafíos tales como movimientos complejos, cambios en la iluminación u oclusiones parciales, entre otros.

Técnicas:

Para lograr extraer dicha información, se han propuesto varios enfoques diferentes a lo largo de los últimos años, incluyendo la **correlación o bloque de coincidencia (*block-matching*), el seguimiento específico de ciertas características (*feature tracking*) y métodos basados en la energía**, donde el método basado en gradiente es uno de los más comunes.

Flujos ópticos (*Optical flows*)

La técnica por excelencia utilizada para la extracción de la información de movimiento se basa en los **flujos ópticos**.

Un flujo óptico se define como el patrón del movimiento aparente de los objetos, superficies y bordes en una escena causado por el movimiento relativo entre un observador y la escena en sí misma. Este flujo proporciona información del movimiento completo en cuanto a la dirección y magnitud del mismo, ya que se basa en vectores. Por tanto, mediante esta técnica es posible definir también la velocidad del objeto que se sigue.
Además, permite realizar una evaluación exhaustiva mediante la división del flujo en dos vertientes. Por un lado, se tendrá el flujo óptico dominante; por otro, el flujo óptico residual, que permitirá obtener el movimiento real del objeto a seguir tras cancelar el movimiento de la imagen dominante que posee información no deseada debido al movimiento de la cámara. La obtención de esta información implicará aplicar estimadores reguladores de energía que preserven las discontinuidades del movimiento.

Una de las técnicas más ampliamente utilizadas para estimar el flujo óptico en un vídeo se basa en el método de **Lucas-Kanade** [16]. Este método supone que el flujo es esencialmente constante en la zona local o vecindario alrededor del píxel en cuestión, resolviendo las ecuaciones básicas de flujo óptico para todos los píxeles en ese vecindario por el criterio de mínimos cuadrados.

Aplicaciones

A la información de movimiento extraída en un noticiario pueden atribuírsele ciertas situaciones típicas que ayuden a diferenciar el contenido que representa.

Estas situaciones pueden hacer referencia a situaciones concretas como el movimiento debido al tráfico en un reportaje de ciudad, o bien ayudar a la identificación de sucesos al combinarse con otras características.

Información de audio

La adicción de la información de audio supone un reto y un esfuerzo extra en la eficiente indexación y recuperación de vídeo.

La mayoría de los sistemas actuales de indexación de audio utilizan reconocimiento de voz. Estos se basan, a grandes rasgos, en obtener transcripciones alineadas en el tiempo para, después, identificar situaciones en el vídeo donde esas transcripciones puedan cobrar sentido [15][17].

Técnicas

Utilizar eficientemente la información de audio en un algoritmo de indexación implicaría seguir las etapas definidas a continuación:

1. Segmentación y clasificación de audio.
2. Análisis del audio extraído para la indexación.
3. Recuperación basada en contenido de audio.

En este ámbito, la discriminación entre habla y música cobra una gran importancia, ya que estas constituyen los dos tipos más amplios de fuentes de audio.

Uno de los métodos más simples estaría basado en la **tasa media de cruces por cero** y las **características derivadas de la energía** extraída en la información, lo que permite establecer umbrales que disciernan entre ambas fuentes.

Otros más complejos tendrían en cuenta los **dominios frecuenciales y cepstrales** y técnicas basadas en **MAP o GMM**. En este ámbito se vuelve más simple la distinción entre habla y música, ya que su distribución espectral y los patrones de cambio temporal son bastante diferentes.

Entre estos métodos, existen aquellos que extraen coeficientes a modo de características, como por ejemplo los coeficientes cepstrales en las frecuencias de Mel (**MFCC**), utilizados para la representación del habla basándose en la percepción auditiva humana.

Aplicaciones

En los noticiarios existen multitud de eventos temporales con información de audio que pueden ser detectados, reconocidos y establecidos como un patrón característico que sea capaz de ayudar en la clasificación de los distintos escenarios.

Para empezar, cualquier noticiario dispone de una sintonía típica que se repite a menudo en el inicio y el fin de la emisión. Además, se establecen cortinillas típicas entre reportajes que puede ser de gran ayuda para segmentarlos.

Además, existen sonidos claramente identificables como el ruido del tráfico, el ruido típico asociado a un evento deportivo, etc. que, combinados adecuadamente con el resto de características hacen robusto al sistema de indexación que se pretende construir.

3.3.1.3 Características espacio-temporales

Podría establecerse este tercer nivel de clasificación, tal y como se hizo en parte de la bibliografía comentada en el Capítulo 2, cuya base estaría en la combinación de características tanto espaciales como temporales. En estos algoritmos se realizará un estudio por zonas, extrayendo volúmenes (información tridimensional) y contrastando con la información temporal extraída.
Dicho de otro modo, se podrían aplicar técnicas que se basaran en la extracción simultánea de características con información relativa a ambos sectores, lo cual implica hacer uso de descriptores no estáticos. Por ejemplo, en [18] se propone un método basado en gradientes en tres dimensiones que proporciona buenos resultados, realizando una comparación con otros ya existentes que persiguen el mismo fin.
Podría establecerse este tercer nivel de clasificación, tal y como se hizo en parte de la bibliografía comentada en el Capítulo 2, cuya base estaría en la combinación de características tanto espaciales como temporales. En estos algoritmos se realizará un estudio por zonas, extrayendo volúmenes (información tridimensional) y contrastando con la información temporal extraída.

Dicho de otro modo, se podrían aplicar técnicas que se basaran en la extracción simultánea de características con información relativa a ambos sectores, lo cual implica hacer uso de descriptores no estáticos. Por ejemplo, en [18] se propone un método basado en gradientes en tres dimensiones que proporciona buenos resultados, realizando una comparación con otros ya existentes que persiguen el mismo fin.

3.4 Conclusiones

Como se puede observar, las fuentes de información que vienen intrínsecas en un vídeo son enormes. A su vez, la información puede ser analizada, extraída y clasificada de acuerdo a múltiples criterios y de muchas maneras diferentes, por lo que la recuperación de un vídeo acorde a sus características puede suponer un gran desafío.

Todas las características anteriormente mencionadas son sólo una pequeña parte de todas las que de un vídeo pueden derivarse, aunque constituyen las más relevantes y comúnmente utilizadas en la tarea que se plantea.

Tendrá, por tanto, que tenerse en cuenta todos estos factores a la hora de construir un sistema de indexación y recuperación de vídeo, enfocándolo al máximo a la aplicación que se quiera desarrollar y teniendo en cuenta el tipo de información que podrá ser útil recuperar (en este caso, tal y como se plantea, enfocada a noticiarios).

Capítulo 4. Diseño y desarrollo del sistema

4.1 Introducción

En este capítulo se detalla el procedimiento elegido y llevado a cabo para crear un **algoritmo de indexación automática de vídeo por medio de la extracción, análisis e indexado de sus características visuales globales**, de modo que, mediante la combinación adecuada de las mismas y tras un análisis previo de las posibles decisiones a tomar, se obtenga un etiquetado sistemático de un vídeo de cara a su posterior recuperación.

De esta manera, se disminuye significativamente el tiempo dedicado a la recuperación de vídeos de interés y permite disponer de una clasificación en repositorios multimedia en función de distintos criterios y de cara a infinitas aplicaciones (académicas, comerciales, ocio/entretenimiento, seguridad, etc.).

En la Figura 7 que se muestra a continuación, se pueden diferenciar las distintas etapas diseñadas e implementadas en el sistema propuesto. Se trata de un método de indexación y recuperación de vídeo basado en la extracción de características visuales globales y que, además, tiene su enfoque principal en vídeos de **noticiarios**, los cuales constituyen una parte importante en bases de datos multimedia.

Hay que tener en cuenta que, el proceso que se lleva a cabo para diseñar y desarrollar el algoritmo de indexación automática de vídeo, está enfocado al conocimiento previo por parte del usuario que realiza la consulta o *query* del contenido que desea encontrar. Del mismo modo, este usuario conoce en cierta manera la generalidad del algoritmo y el nivel de detalle al que tiene que someter su consulta. Se plantea un caso a modo de ejemplo que matiza esta aclaración:

"Se propone el caso particular de un usuario que pretende buscar una noticia relacionada con una inundación que ha ocurrido recientemente en su comunidad autónoma y que ha dejado a la ciudad con graves daños por el desastre. Así pues, este usuario desea encontrar un evento que se caracteriza por los siguientes aspectos: evento localizado en un entorno de exterior; además, se trata de una zona urbana y donde el movimiento de la imagen será alto debido a la fuerza del agua que circula por las calles, entre otras. Haciendo una búsqueda con estos términos ("exterior", "urbano", "alto movimiento") en una base de datos donde el usuario ya conoce que se encuentra este material de vídeo que desea encontrar, disminuirá la cantidad de vídeos recuperados tras la consulta, de modo que acotará más la búsqueda y podrá realizarla de forma más sencilla, rápida y eficiente."

Esto es así ya que el fin que se persigue con la implementación de este algoritmo es poder **simplificar la búsqueda de un vídeo** en concreto o de un material multimedia en concreto. Si, además, luego se combina esta búsqueda con información adicional, como es información relativa al audio, la eficiencia en la recuperación del material deseado aumenta. Continuando con el caso propuesto como ejemplo, podría aunarse estas características visuales del vídeo con la identificación de sonido de agua, lluvia o temporal de viento, realizando así un mayor filtrado de información.

Este último punto, la inclusión de información relativa al audio, no se tiene en cuenta para este trabajo y se plantea como una posible vía para el trabajo futuro, el cual se comentará más en profundidad en el Capítulo 6.

Hay que tener en cuenta que, no siempre el uso de un número mayor de características va a ser capaz de obtener mejores resultados. En general, cuanto mayor es la información de la que se dispone, mayor es también el tratamiento que habrá que dar a dicha información, analizándola y procesándola adecuadamente para poder derivar en mejores resultados.

Existen casos excepcionales, como lo es el caso en que los nuevos datos introducidos tienen una alta correlación con los ya disponibles, en el que el sistema sí puede experimentar una cierta mejora, pero también está condicionado en gran medida por la capacidad de discriminación o detección de la información que tenga el sistema.

Este capítulo se estructura en diferentes secciones: en primer lugar, se hará un resumen de los requisitos que se pretenden cubrir con el sistema creado para, a continuación, comentar en mayor profundidad las etapas seguidas y los módulos desarrollados, describiendo las técnicas utilizadas para conseguir cumplir con dichos requerimientos.

4.2 Requerimientos iniciales

Los **requerimientos** que se establecen a priori y cuya consecución se persigue durante el diseño y desarrollo del sistema son los citados a continuación, siempre buscando un método general como primera aproximación a un sistema robusto de indexación:

- **Completo:**

Se busca crear, partiendo de cero, un sistema que contemple toda la cadena de procesamiento, buscando dar la mejor respuesta de acuerdo a los criterios establecidos.

- **Simple:**

Se pretende que un usuario cualquiera, no obligatoriamente formado en el ámbito del tratamiento y el procesamiento de vídeo, pueda ser capaz de intuir el funcionamiento del sistema y de utilizar los recursos que se le proporcionen para recuperar vídeos con el contenido deseado.

- **Parametrizable (adaptable):**

Una de las principales ventajas de las que dispondrá el sistema propuesto es su capacidad para adaptarse a los cambios que vayan considerándose necesarios o que, o bien, quieran irse moldeando de una u otra manera en función del alcance que se pretenda obtener.

- **Robusto y eficiente:**

Siendo conscientes del alcance del que, en primera instancia, se dispone con este sistema, se pretende que el algoritmo sea lo suficientemente sólido como para arrojar resultados considerablemente acertados para un conjunto de vídeos; siendo válida la posible adaptación del entorno a ciertos parámetros para conseguirlo.

- **Coste *(low cost)*:**

Se persigue un coste medio-bajo a nivel computacional, que no implique excesivos tiempos de ejecución y rendimiento en la máquina.

4.3 *Overview*

A continuación se muestra un esquema con los bloques principales de los que consta el algoritmo propio y se reflejan en él las distintas etapas establecidas en el diseño y en la implementación:

Antes de entrar en detalle con lo desarrollado en cada módulo, se hace un repaso rápido sobre el esquema representado en la Figura 7.

En primer lugar, se dispondrá de una base de datos con los vídeos que se utilizarán para ejecutar y evaluar el sistema. Este conjunto de vídeos serán utilizados mientras se desarrolla el algoritmo, donde se irá probando progresivamente la adaptación del sistema a los requerimientos establecidos a priori. Estos vídeos están extraídos de noticiarios o programas informativos, combinando reportajes muy variados y escenas de múltiples características.

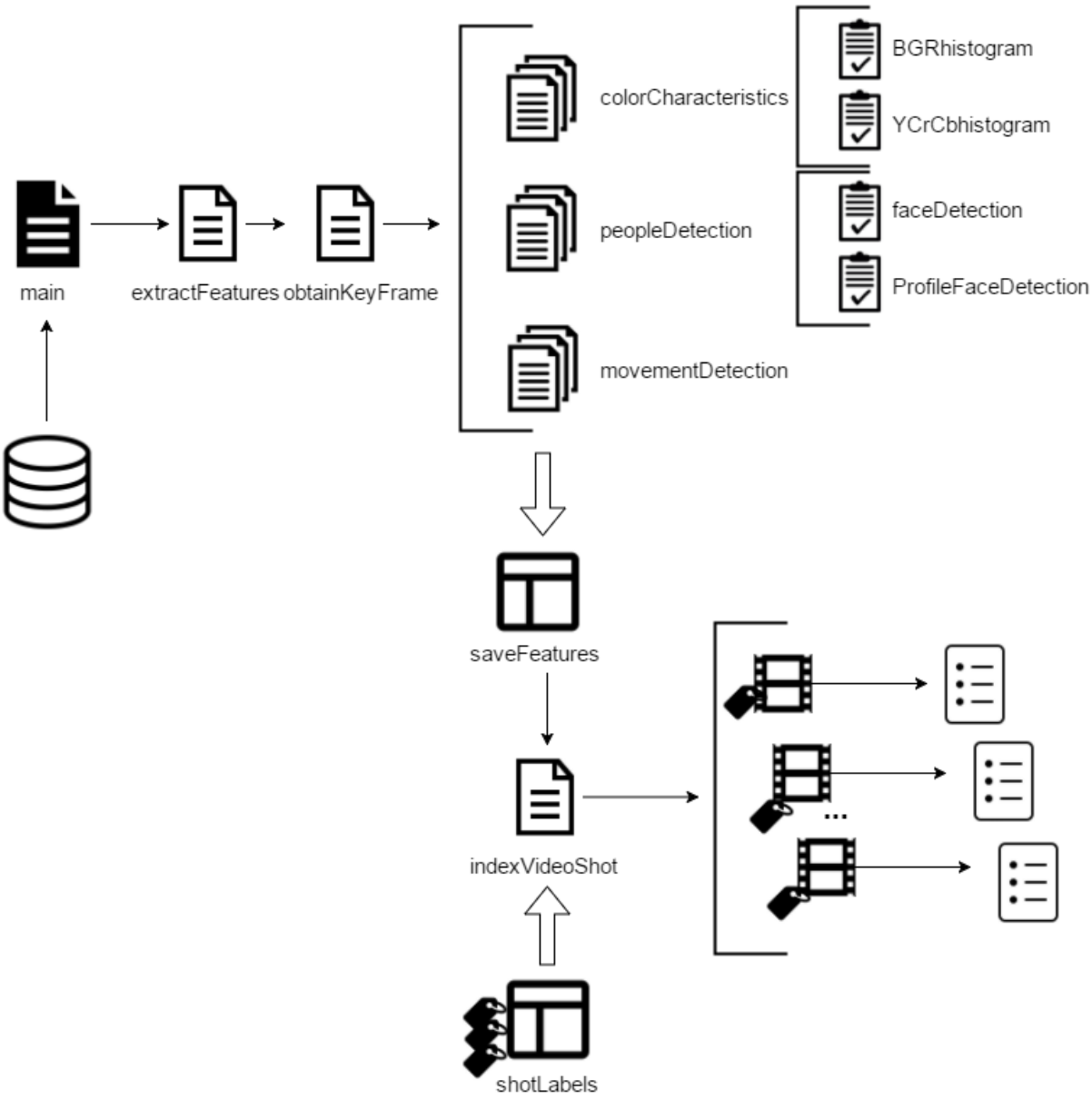

Figura 7. Esquema del algoritmo de indexación propio. Fuente: Propia

Figura 8. Imágenes capturadas de los vídeos de noticiarios utilizados. Fuente: Propia

En el bloque principal o *main*, se estructura el flujo del algoritmo. En este módulo se leerán cada uno de los vídeos y serán enviados al siguiente bloque, donde se extraerán las características. Dependiendo del tipo de característica, será necesario acudir a un nivel distinto de los mencionados en la jerarquía de estructuración de un vídeo, normalmente bien a nivel de *shot* o bien a nivel de *key frame*, esta última obtenida en el módulo *obtainKeyFrame*.

Las características serán extraídas en uno u otro bloque en función del tipo que se trate (extracción del color, detección de personas, detección de movimiento) y, tras ello, se almacenarán en una clase, *saveFeatures*, para posteriormente ser capaces de recuperarlas e indexar o etiquetar los *shots* en *indexVideoShot*. Estas etiquetas estarán definidas en la clase *shotLabels*.

Tras haber establecido de manera breve los objetivos a perseguir con el sistema propuesto, se desglosan a continuación las distintas etapas y módulos ya mencionadas en los que se ha dividido el sistema y cuyo diseño e implementación ha ocurrido de forma secuencial, asegurando el buen funcionamiento de las partes para obtener la actuación general deseada finalmente.

4.4 Segmentación temporal

A continuación se detallan las etapas seguidas para realizar la segmentación del algoritmo. Esta segmentación está basada en una aproximación sencilla, como se verá a continuación, pero que funciona bien con el *dataset* que se dispone y que además implica una baja carga computacional, por lo que resulta útil y cómodo para el fin que se persigue.

4.4.1 Cálculo de disparidad entre *frames* consecutivos

El primer paso se centra en la segmentación de los vídeos utilizados como *dataset* de desarrollo, de modo que puedan clasificarse posteriormente de acuerdo a los criterios que se consideren, los cuales se comentarán más adelante.

La técnica que se implementa para la segmentación frente-fondo se basa en el ***frame differencing*** [21].

El cambio debido al movimiento que se produce entre sucesivas imágenes contiene mucha información. Esta información puede ser extraída mediante el procesado de las imágenes de vídeo.
La complejidad en esta técnica reside en mantener una buena calidad a pesar de que múltiples factores puedan afectarla, como el entorno o ambiente que existe durante el proceso de adquisición de las imágenes. En este entorno, se debe tener en cuenta características como los cambios de luminosidad, las condiciones climáticas que pueden ser variables, la cantidad de objetos en movimiento en la misma escena, etc.
Una de las técnicas más comúnmente utilizadas para la segmentación de fondo en el procesado de vídeo es la técnica de *frame differencing*, debido a su baja complejidad y a su buen funcionamiento, especialmente cuando se combina con otros métodos que refuerzan su actuación. Su procedimiento se basa en **detectar los cambios entre dos *frames* adyacentes**.

Suponiendo que un *frame* en un momento *t* se define como $f(x,y,t)$ y que la subsecuente frame se define como $f(x,y,t+1)$, la operación de *frame differencing* que resulta en una imagen binaria sería la siguiente:

$$D(x,y,t+1) = \begin{cases} 1, & |f(x,y,t) - f(x,y,t+1) > Th| \\ 0, & resto \end{cases}$$

Donde *Th* es el umbral fijado previamente. Este valor que se le da al umbral puede obtenerse de diferentes maneras, existiendo técnicas de ajuste automático y adaptativo. Otra vía posible

es fijarlo de manera experimental al observar el vídeo a segmentar y decidiendo así un valor óptimo para la correcta y precisa segmentación.

El algoritmo implementado leerá los sucesivos *frames* de un vídeo, convirtiéndolos a escala de grises y realizando la resta o diferencia de los dos *frames* adyacentes, generando una imagen diferencia de forma consecutiva. Cuando se ha generado esta imagen, se fija el umbral (mediante la experimentación) y, cuando la variación entre estos *frames* supera el umbral establecido, se segmentará el vídeo indicándose cambio de *shot*.

4.4.2 Obtención de *shots*

A continuación y, una vez se ha realizado el cálculo de disparidad entre *frames* en una primera aproximación, se procede a ascender un nivel más en la jerarquía de descomposición de un vídeo, como se observaba en la Figura 2 del Estado del arte, hasta diferenciar distintos *shots*, a partir de los cuales extraer características y clasificar de forma más general un vídeo.

Esto se decide hacer así dado el tipo de vídeo que se está utilizando como *dataset*. Es decir, al disponer como conjunto de prueba vídeos extraídos de noticiarios, se podrán definir *shots* durante su visualización, estando cada uno relacionado con un tipo de contenido usual en este registro (reportaje, plano de estudio, etc.).

Concretamente, una noticia se divide en diferentes segmentos. Cada segmento está dividido en *shots* diferenciados y que, en función de la categoría identificada (reportaje, plano de estudio, etc.) pueden agruparse en *monoshot* o *multishot*, siendo este último el que se relaciona con los reportajes. Es decir, en un reportaje pueden aparecer distintos shots que, en realidad, pertenecen a un mismo tema o que tratan sobre un mismo contenido.
Por tanto, se podría establecer una taxonomía en este registro (los noticiarios) como la que se muestra en la siguiente figura:

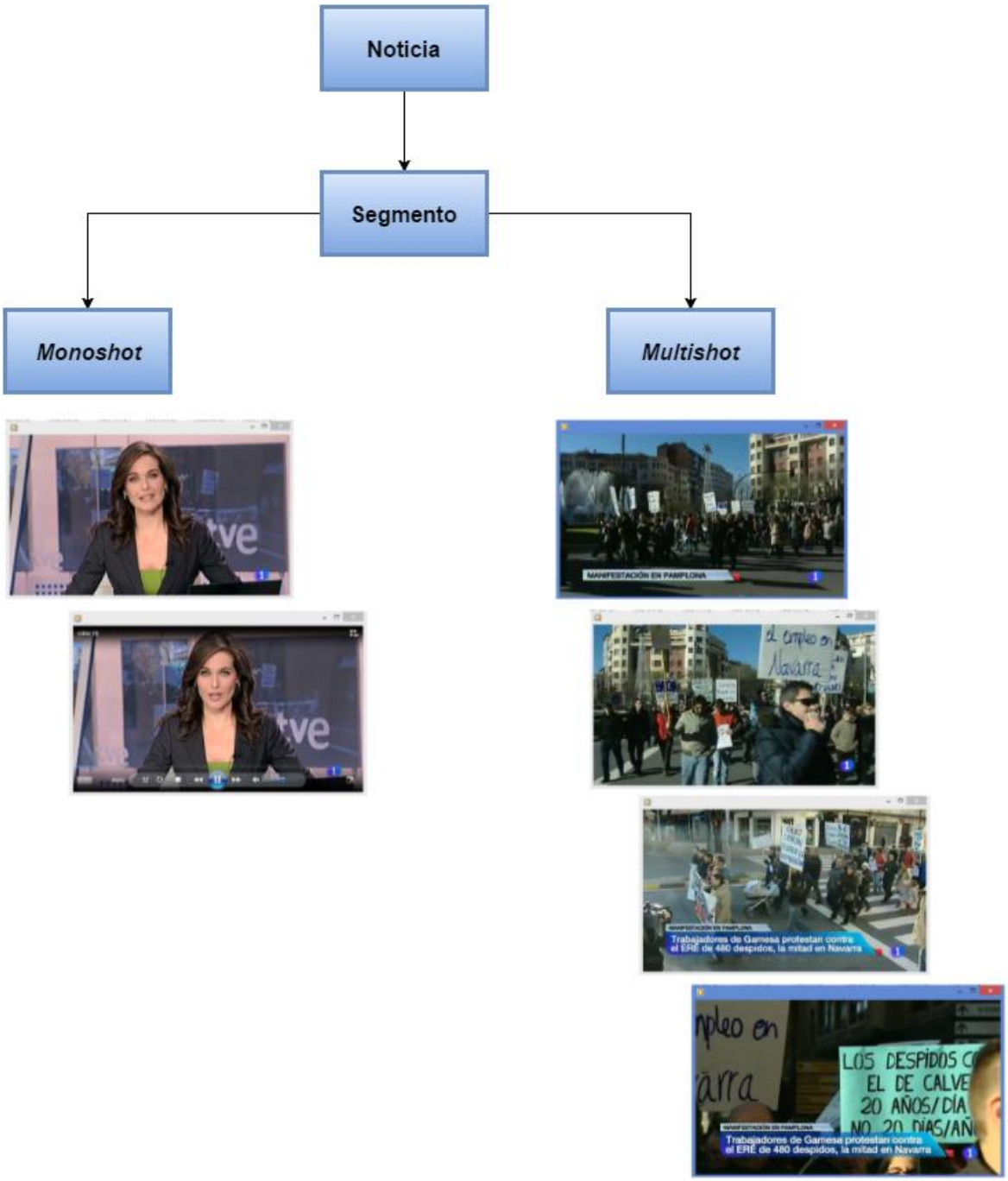

Figura 9. Taxonomía identificada en noticiarios Fuente: Propia

El procedimiento a seguir para obtener los distintos *shots* se basará en observar cuándo **la variación en la imagen diferencia supera el umbral establecido**, lo que llevará implícito un **salto más significativo en la imagen** (por ejemplo, cambio de la toma del presentador en estudio dando paso a la noticia al reportaje en sí mismo). En cada uno de estos saltos se diferenciará un *shot*.

El **umbral** será fijado a partir de un estudio previo con distintos valores (más del 50% de la imagen ha sufrido variación, más del 60%, más del 70%...) hasta obtener el más discriminante. Con este valor se decidirá que el cambio en la imagen es lo suficientemente grande como para considerarlo un nuevo *shot*.
Una consideración importante que se debe tener en cuenta es que, si en la imagen se produce mucho movimiento, la diferencia entre una *frame* y la siguiente puede aumentar de tal modo que el algoritmo lo considere, de forma errónea, cambio de *shot*. Este matiz debe introducirse como un parámetro adicional en el que indicar la longitud mínima admitida de un *shot*, teniendo en cuenta que el vídeo ha sido grabado a 25 fps.

Finalmente, se decide tras diversas pruebas con los *dataset* utilizados que se realizará la segmentación estableciendo el salto como cambio de *shot* cuando la variación en la imagen diferencia es mayor del 50% y estableciendo, de igual modo, que como mínimo un *shot* puede durar 1 segundo (25 fps).

4.4.3 Selección de key frames

Una vez se obtienen los diferentes *shots*, habrá que seleccionar un *frame* representativo del *shot* en cuestión, al que llamaremos *key frame* y que será el *frame* central del *shot*. A partir de él, se procederá a la extracción de las características que nos permitirán clasificar el vídeo bajo estudio. Esta aproximación se ha tomado por varias razones, como la simplificación del algoritmo (lo que conlleva una reducción en la carga computacional necesaria), siendo una solución que se ajusta a las necesidades que se presentan.

A continuación se muestra en la Figura 8 las diferentes etapas que pueden observarse en la segmentación, realizada sobre uno de los vídeos del *dataset*.

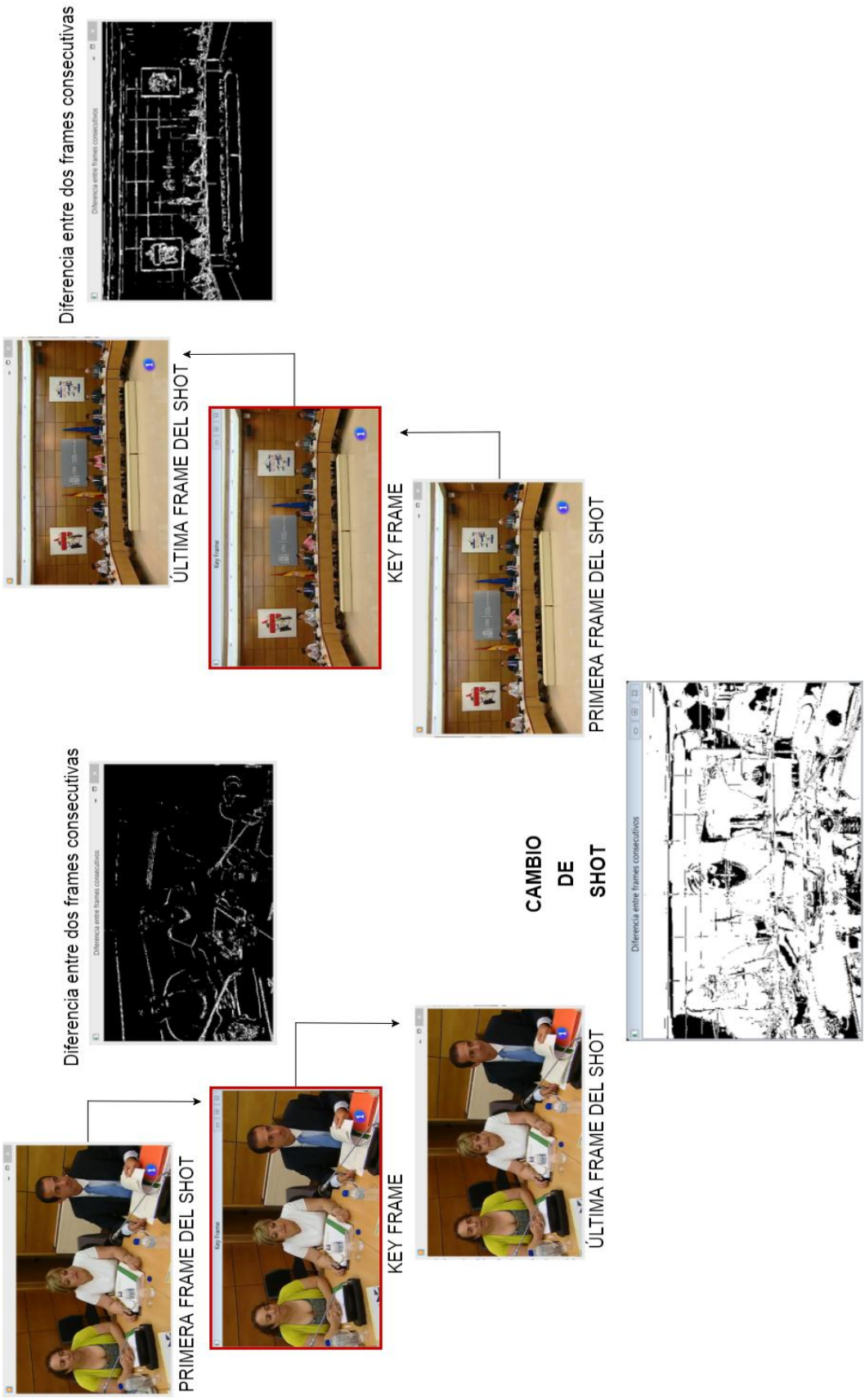

Figura 10. Ejemplo de segmentación de *shots*. Fuente: Propia

4.5 Indexación

4.5.1 Extracción de características

Se procede a continuación a extraer una serie de características en los distintos vídeos utilizados como *dataset* para probar el algoritmo que va a ser implementado. Este *dataset* se compone, tal y como se ha descrito, de distintos vídeos específicos de noticiarios, los cuales aúnan escenas de distinta índole, combinadas en diferentes entornos (escenas de interior, de exterior, planos más o menos alejados, etc.).

El procedimiento de extracción de características se realiza en un bucle o ciclo cerrado donde se siguen los siguientes pasos:

1. Cada vez que uno de los vídeos es leído y dividido en sus múltiples *shots*, como ya se ha comentado, cada *shot* quedará representado por su *key frame*.

2. A partir de esta *frame* clave, se extraerán una serie de características de forma consecutiva, al pasar durante la ejecución por distintos módulos del algoritmo, cada uno centrado en la extracción de una característica en concreto.
   Por ejemplo, uno de los módulos tendrá como objetivo la extracción de la característica en relación al color, la luminancia y la crominancia de la imagen, obteniendo valores particulares para la misma (y por consiguiente, para el *shot* bajo análisis).
   Es importante resaltar que para características que poseen información temporal la extracción se hará a nivel de *shot*, entendiendo éste como la máxima granularidad posible en este aspecto.

3. A medida que las características van siendo extraídas, se guardan en una clase llamada *saveFeatures;* una estructura que almacena en distintos campos los valores para cada una de las características que han sido reconocidas en la *key frame* o en el *shot* y que, más adelante y tras el análisis de las mismas, servirán para etiquetar al vídeo en cuestión.

Estas tres etapas consecutivas se repiten durante todo el proceso de análisis y extracción de características.

En el capítulo 2, el estado del arte, se proponía la clasificación en base a la división, en primera instancia, entre características espaciales y temporales. Aunque esto puede ser así, es importante señalar que existen otras muchas clasificaciones en función del objetivo de la

aplicación, donde una muy común se basa en distinguir desde un primer momento entre características estáticas y dinámicas, y es en base a esto como se divide nuestro algoritmo.

4.5.1.1 Módulos implementados para la extracción

Se detallan ahora uno a uno los módulos de la función ***extractFeatures,*** encargada de abstraerse de la imagen reconocida como *key frame* (o del *shot*, en caso de características temporales) y de obtener a partir de ellos una serie de características globales. Estos módulos implementados son los siguientes:

colorCharacteristics

A través de la información de color puede obtenerse un método eficaz para clasificar y, posteriormente, indexar el contenido de la imagen. Al ser el color una característica estática y espacial, se utilizará la *key frame* como punto de partida para la extracción.
Esta información de color puede extraerse mediante diferentes descriptores/características de color, aunque dos de las más significativas y útiles para el fin que perseguimos y que, serán las implementadas en nuestro algoritmo, son las descritas a continuación:

BGRhistogram:

El espacio de color RGB (en el entorno de desarrollo utilizado, OpenCV, se estructura como BGR) hace referencia a los tres canales de color clásicos en una imagen: azul (*blue*), verde (*green*) y rojo (*red*).

YCrCbhistogram:

En este otro espacio de color, los canales en los que se divide hacen referencia a la luminancia y crominancia (diferenciadas en dos tipos, Cr y Cb).

En ambas funciones, se calculará el histograma para cada canal del espacio de color y se realizará el cálculo del porcentaje que ocupa cada uno de ellos en la imagen. Esto se hará en **dos vías** diferentes, según las cuales se estimará este porcentaje de cada canal (ya sea para BGR o para YCrCb):

1. **Global** u **holística**.
2. **Por partes** o **local**.

Definiendo el procedimiento seguido para cada una de ellas, se tiene:

1. La extracción de la característica del color **de forma global u holística** se basa en tomar a la *key frame* de cada *shot* al completo y en ella realizar el histograma y calcular los porcentajes de representación de cada canal como resultado.
   Esta aproximación es la que se sigue en una primera instancia pero, finalmente, el procedimiento que se aplicará y será el definitivo es el que se comenta en el siguiente punto.

2. Al realizar la división de la *key frame* en **bloques o *patches (de forma local)***, se profundiza en mayor medida en el análisis de cada zona en la que se ha dividido la imagen.
   Se utiliza en este caso un *grid* de 3x3 celdas, por lo que la imagen de análisis se queda dividida en 9 subimágenes donde analizar la contribución de cada canal. El tamaño de subdivisión dado ha sido elegido al considerar el tamaño de la imagen total y el de las partes en las que queda dividida tras aplicar dicho *grid*, lo cual se cree razonable para profundizar más en el análisis del color. Así, se obtiene para cada una de las partes un valor diferente y más específico, decidiendo finalmente el que más ocurrencias tiene en la imagen total.
   Con esta aproximación, se entra más en detalle en cada zona, obteniendo un valor más ajustado y más realista del atributo que se analiza (en este caso, el color).

En la siguiente Figura 10, se muestra un *frame* arbitrario junto con parte de los resultados que se muestran por pantalla durante la ejecución. Concretamente aparecen, señalados en amarillo, los valores de cada espacio de canal atribuidos a la primera subdivisión en la que queda diferenciada la imagen tras aplicar el *grid* 3x3 comentado.

Primeramente se da la media para cada canal, procesados en ambos espacios RGB o YCrCb, para después calcular el porcentaje que ocupa dicho canal en esa subdivisión. De este modo, se decide cuál es el canal dominante en ella.

Por ejemplo, en el caso que se presenta, se decide que es el rojo el canal con más presencia, al igual que el valor para las crominancias es significativamente mayor. Observando el *frame*, se ve que los valores que se devuelven tienen sentido:

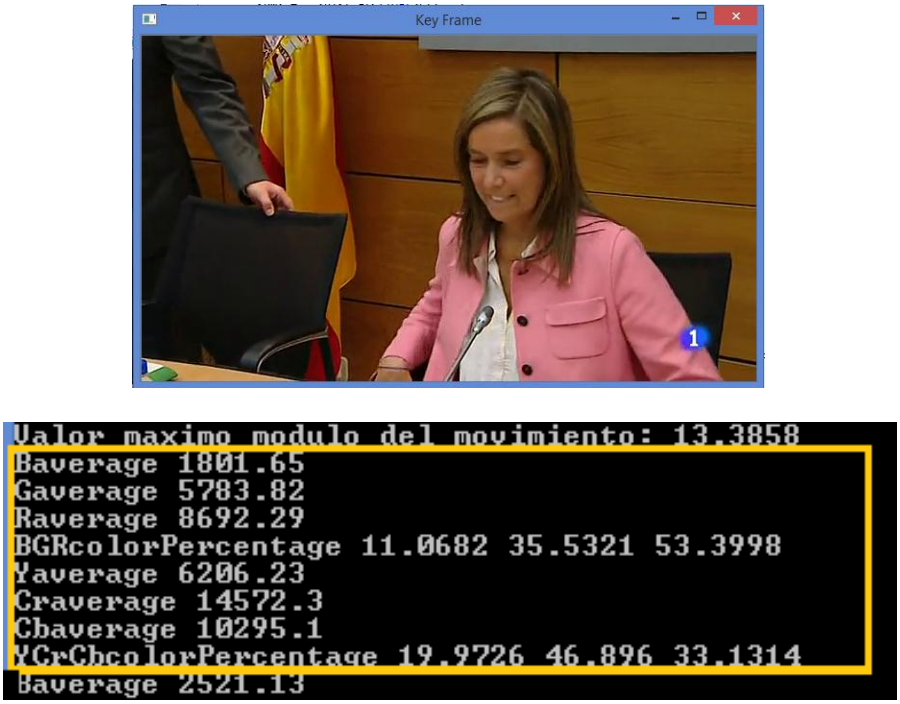

Figura 11. Representación de una *key frame* arbitraria y sus correspondientes valores por canal en cada espacio analizado (RGB y YCrCb.) para la primera subdivisión tras aplicar el grid 3x3 Fuente: Propia

peopleDetection

Otro de los bloques más representativos del sistema de cara a poder indexar en distintas categorías relacionadas con partes de un noticiario es aquel que se encarga de la detección de personas.

Gracias a las distintas funciones proporcionadas en la biblioteca de OpenCV esta tarea puede llevarse a cabo. Mediante el uso de dichas funciones, es posible detectar personas en un vídeo, pudiendo estar definidas y acotadas desde muchos criterios diferentes: figura total de la persona, parte superior del cuerpo (tronco, cabeza y extremidades superiores), caras desde una perspectiva frontal, caras de perfil, etc. Se trata de un análisis muy amplio y que puede tener muchos análisis y contextos de aplicación diferentes, por lo que el enfoque dado vendrá determinado en gran medida por el alcance.

El primer paso en este módulo se basa en una etapa de entrenamiento con diferentes plantillas (de entre las proporcionadas por la herramienta se eligen aquellas que puedan adaptarse mejor a las necesidades del sistema).

Tras sopesar las diferentes opciones y realizar varias pruebas tras el ajuste de los parámetros pertinentes, se decide centrar la detección de personas en la **detección de caras**, tanto desde una **perspectiva frontal** como desde una **perspectiva de perfil**.

Esto es así ya que, si se utiliza una plantilla que pretenda detectar la figura completa de una persona, en conjunto no se van a obtener buenos resultados, debido a que en la mayoría de los casos con los vídeos que se trabaja; es decir, con el *dataset* concreto que se utiliza, las personas aparecen de medio cuerpo (segundos planos en tomas de estudio, por ejemplo) o primeros planos (entrevistas en los reportajes). Se distinguen por tanto dos módulos:

hogFaceDetection:

Enfocado a detectar las caras desde una **perspectiva frontal**. La plantilla utilizada es *haarcascade_frontalface_alt.xml*.

El algoritmo utiliza una **detección multiescala** y una **clasificación en cascada basada en Haar**, lo cual hace referencia a que el clasificador se compone de varios clasificadores sencillos (etapas) que se aplican a una región de interés de forma consecutiva hasta que en algún momento se rechaza el candidato o se pasan todas las etapas. Además, la función *face_cascade.detectMultiScale* requiere **ajustar los parámetros** de entrada, como los pesos que se dan a las caras para detectarlas como tal o, en caso contrario, descartarlas. Este ajuste se ha realizado en sucesivas pruebas observando qué valores, en general, se adaptaban mejor al *dataset* utilizado.

Figura 12. Detección de caras frontales en los vídeos de *training*. Fuente: Propia

hogProfileFaceDetection:

En este caso, enfocado a detectar las caras desde una **perspectiva de perfil**. El procedimiento es exactamente el mismo que el detallado para las caras frontales, salvo con la excepción de que la plantilla utilizada será otra, haarcascade_profileface.xml, y los parámetros deberán ser sometidos a otro ajuste que se adapte adecuadamente a la nueva detección. Este ajuste será realizado del mismo modo que se hizo en el caso de las caras frontales, mediante los *dataset* utilizados y su experimentación,

Se obtendrán tras este proceso dos vectores, uno la posición y dimensión de todas las caras frontales detectadas en la imagen y otro idéntico para el caso de las caras de perfil. Igualmente, de estos vectores puede obtenerse el número de caras presentes en la imagen, parámetro que será el utilizado para su posterior análisis en la indexación.

movementDetection

El movimiento presente en un vídeo contiene mucha información y puede ser de gran ayuda para ciertos casos en la indexación, así como para otras múltiples aplicaciones como el seguimiento de objetos (*tracking*), para el que además de la cantidad de movimiento se necesita disponer de información sobre la trayectoria del mismo.

Una forma muy común de detectar el movimiento en una imagen se basa en la implementación de un **flujo óptico,** como ya se comentó en el Capítulo 3. Un flujo óptico describe el patrón aparente de movimiento en los objetos de una imagen durante *frames* consecutivos. Se suele representar mediante vectores 2D, donde se muestra el desplazamiento que se produce en la imagen siguiendo los puntos que recorre en los sucesivos *frames*. Además, este algoritmo asume dos cuestiones:

1. La intensidad de los pixeles de un objeto no cambia entre *frames* consecutivos.
2. Los píxeles vecinos poseen un movimiento similar.

En nuestro algoritmo, se hace uso de la función propia de OpenCV calcOpticalFlowFarneback.

En ella se sigue el método inicial descrito por *Lukas-Kanade [16]* y que más tarde deriva en el algoritmo de *Gunnar Farnerbäck [20]*, el cual localiza el flujo descrito por cada píxel de la imagen previa del siguiente modo:

$$imagen\ previa\ (y,x) \approx imagen\ siguiente(y + flujo(y,x)[1], x + flujo(y,x)[0])$$

De igual modo que ocurría con las funciones de detección de personas, mediante la experimentación con el *dataset* de estudio, se ajustarán ciertos parámetros para conseguir adaptar la detección en nuestro sistema, entre los que destaca el tamaño de la ventana en que se detecta movimiento. Este tamaño de ventana supone un compromiso entre la precisión del movimiento que quiera detectarse (en ese caso, el tamaño de ventana será menor) y la capacidad que tendrá el algoritmo para detectar movimientos de mayor magnitud (en cuyo caso, el tamaño será mayor).

Una vez se ha ejecutado esta función se pasará a analizar el movimiento detectado mediante tres aspectos:

- **Módulo o cantidad de movimiento:** Obtenido el movimiento descrito por los píxeles en los sucesivos *frames*, se procederá a calcular el módulo de los vectores de desplazamiento generados. De este modo, a mayor módulo, mayor desplazamiento.
- **Media de la cantidad de movimiento:** A partir de la imagen generada con todos los módulos de movimiento de cada píxel, se obtiene la media de éstos, guardando un valor orientativo de movimiento para cada *shot*.
  Se estudia en el siguiente apartado, lo que puede considerarse como alto o bajo movimiento en base a dos parámetros (longitud del *shot* mínima aceptada y umbral de la media de los valores de píxel en la imagen diferencia), para su futura indexación.
- **Valor mínimo y máximo de movimiento en cada *shot*.**

4.5.1.2 Módulos implementados para la indexación

En este apartado se describe el procedimiento seguido para indexar a nivel de *shot*, a partir de las características que han sido extraídas y guardadas en la clase *saveFeatures*. Es importante destacar que la indexación se realiza a nivel de *shot*, principalmente por la siguiente cuestión:

- El número de situaciones, escenarios y contextos diferentes que aparecen en un mismo informativo o noticiario es muy elevado, por lo que la indexación a nivel del vídeo completo no tiene demasiado interés y no resulta efectiva para el caso que se propone, ya que se pretende recuperar contenidos con características más concretas.

Analizamos cada módulo de indexación (divididos según la característica a indexar) a continuación:

indexByColor

A partir de los vectores con los porcentajes atribuidos a cada canal para cada uno de los *key frames* representativos de cada vector, podemos obtener de forma cualitativa el canal dominante en los mismos.

indexBGR:

Para el caso del espacio de color BGR, se definirán tres etiquetas:

BLUE	GREEN	RED

indexYCrCb:

Para el caso del espacio de color YCrCb, se definirán dos etiquetas:

LUMINANCE	CHROMINANCE

A partir de ellas pueden deducirse aspectos del *shot* bajo análisis. Por ejemplo, puede llegar a discernirse escenas tomadas de día de aquellas tomadas de noche, si se corresponden a luz natural (exteriores) o artificial (interiores), etc. Se verán en el siguiente capítulo las asociaciones seguidas entre ellas para diferenciar una u otra etiqueta.

indexpeopleDetection

Para indexar en función de las personas que se detectan en la imagen se va a tener en cuenta dos factores diferentes:

1. **Número de personas en la imagen**: Detectará si hay una o varias personas en el *key frame* que se analiza, dando el siguiente valor de etiqueta:

SINGLE PERSON	MULTIPLE PEOPLE

2. **Tipo de plano con que se graba a dicha persona:** En función de la lejanía con la que la persona aparece en la imagen se diferenciarán dos casos, primer plano o plano general:

CLOSEUP	BACKGROUND

Para implementar este segundo caso, se realiza un estudio previo que establezca, de manera orientativa, cuál sería el tamaño genérico de una cara en comparación con las dimensiones de un *frame* cualquiera. De esta manera, se decidirá que la toma se reconoce como una de primer plano.

Se considera por tanto, tras observar varios casos experimentales, que una **relación de aspecto mayor de 1/16 respecto al tamaño del *frame* total** es una buena aproximación de primer plano. Es decir, cualquier cara detectada cuya dimensión sea mayor que un 1/16 del resto de la imagen, se considerará como primer plano.

indexmovementDetection

El análisis para la indexación por movimiento requiere un análisis más profundo y la consideración de otros aspectos importantes que se mencionan a continuación.

Sabiendo que los vídeos del conjunto de *datasets* utilizados han sido grabados a 25 fps, analizamos el movimiento en cada *shot* para establecer un umbral que discierna entre alto o bajo movimiento, en base a dos variables: la **longitud de la duración mínima permitida** de un shot y el **umbral** con el que, al superarlo, se decide nuevo *shot* tras obtener la media de los píxeles de la imagen diferencia entre dos *frames* consecutivos.

Los casos que se plantean son los siguientes:

- **th = 127.5 (50% de la imagen), long_mín = 25 frames (1 seg)**

Se establece cambio de *shot* cuando existe una variación mayor del 50% en la media del valor de los píxeles en la imagen segmentada resultante de la diferencia entre *frames* consecutivas; a la vez que se considera que un *shot* no va a poder tener una duración menor a un segundo.

- **th = 127.5 (50% de la imagen), long_mín = 50 frames (2 seg)**

Con el mismo umbral en la variación de la imagen diferencia, se contempla en este caso que un *shot* no puede tener una duración inferior a 2 segundos. Este es el caso menos exigente.

- **th = 127.5 (50% de la imagen), long_mín = 12 frames (≈ 0.5 seg)**

De nuevo con el mismo umbral, se relaja la condición a una duración mínima de 0.5 segundos para el *shot*, lo que hace que el algoritmo tenga que ser mucho más exigente para no detectar *shot* de manera errónea (por cambios bruscos de movimiento, por ejemplo).

Del mismo modo, se repite este mismo procedimiento para los distintos valores de **long_mín (1, 2 y 0.5 seg)** pero ahora haciendo que el umbral a superar para considerar nuevo *shot*

necesite disponer del **60%** de píxeles con variación en la imagen diferencia entre *frames* consecutivas.

Una vez ha sido realizada esta prueba, se registran los valores de la media de movimiento en la imagen para cada vídeo y para cada uno de sus *shots*, de modo que se generan tras la ejecución (y por consiguiente, análisis) de cada vídeo, un registro de valores, que se muestran en el eje de abscisas en las sucesivas figuras.

Se decide hacer el estudio con un conjunto de 10 vídeos y, tras la visualización y comparación de los resultados en los distintos casos analizados, se establece un umbral para cada caso que diferencie de manera razonable entre bajo o alto movimiento.

Tras establecer dicho umbral particular para cada análisis, se realiza la media de dichos umbrales y se establece el mismo como **umbral genérico de decisión para la detección del movimiento.** Por encima de este umbral se detectará alto movimiento, quedando bajo movimiento cuando no se supera su valor:

LOW MOVEMENT	HIGH MOVEMENT

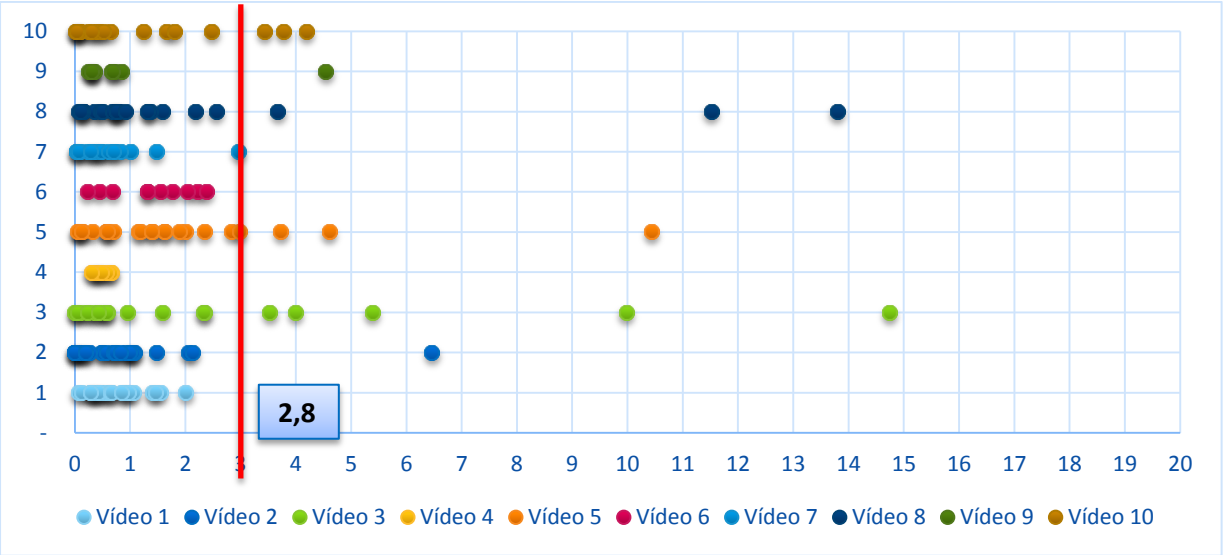

Figura 13. Detección de *shots* con th = 127.5 (50% de la imagen) y *long_mín* = 25 *frames* (1 seg). Fuente: Propia

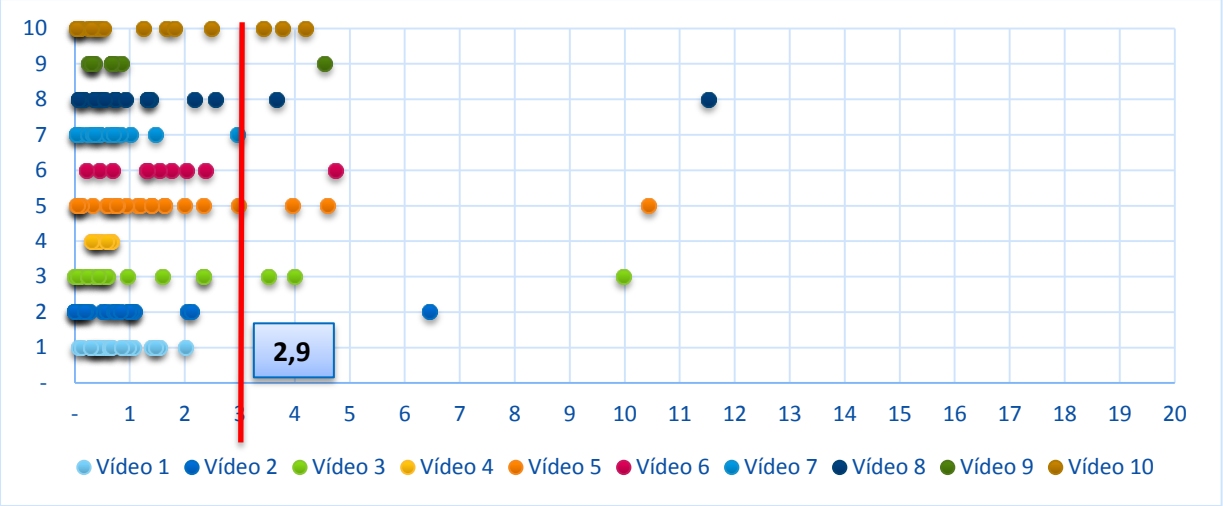

Figura 14. Detección de *shots* con th = 127.5 (50% de la imagen) y *long_mín* = 50 *frames* (2 seg).

Fuente:

Propia

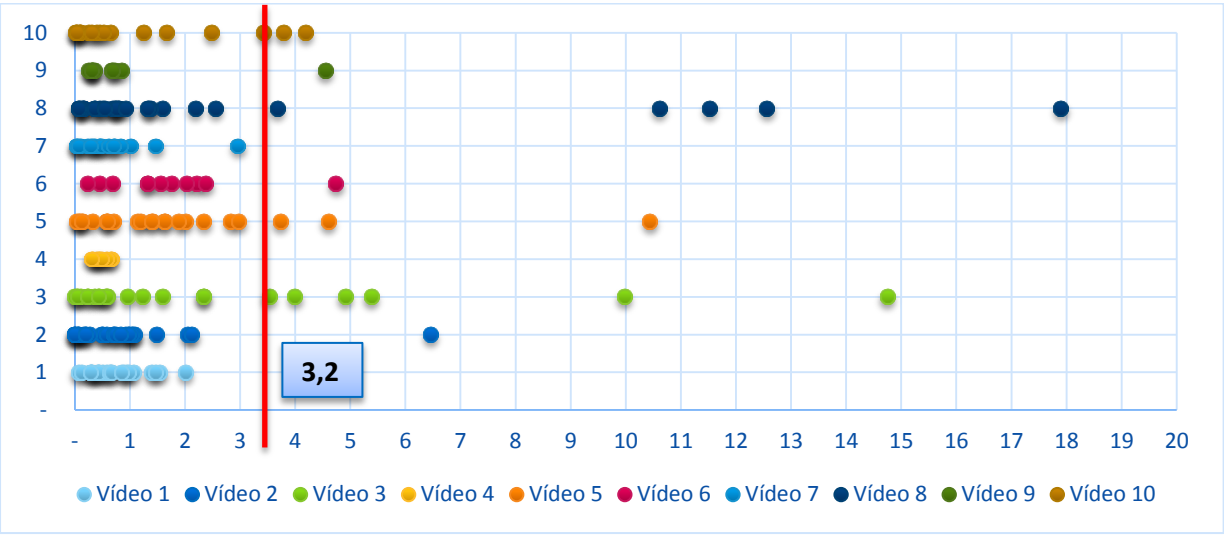

Figura 15. Detección de *shots* con th = 127.5 (50% de la imagen) y *long_mín* = 12 *frames* (≈ 0.5 seg).

Fuente: Propia

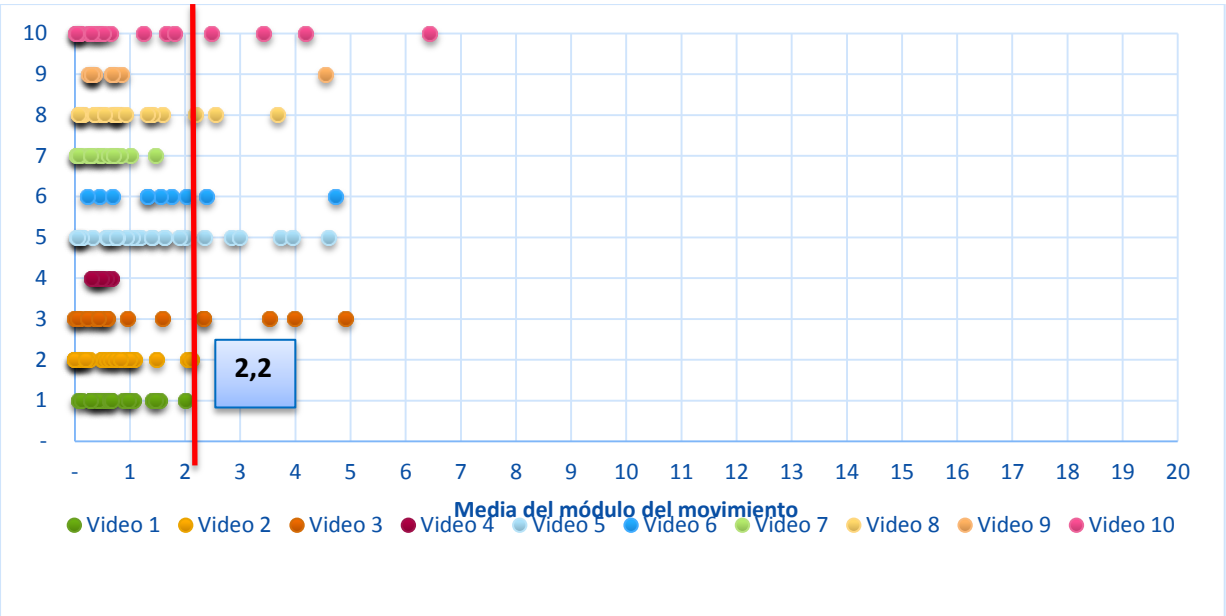

Figura 16. Detección de *shots* con th = 153 (60% de la imagen) y *long_mín* = 25 *frames* (1 seg).
Fuente: Propia

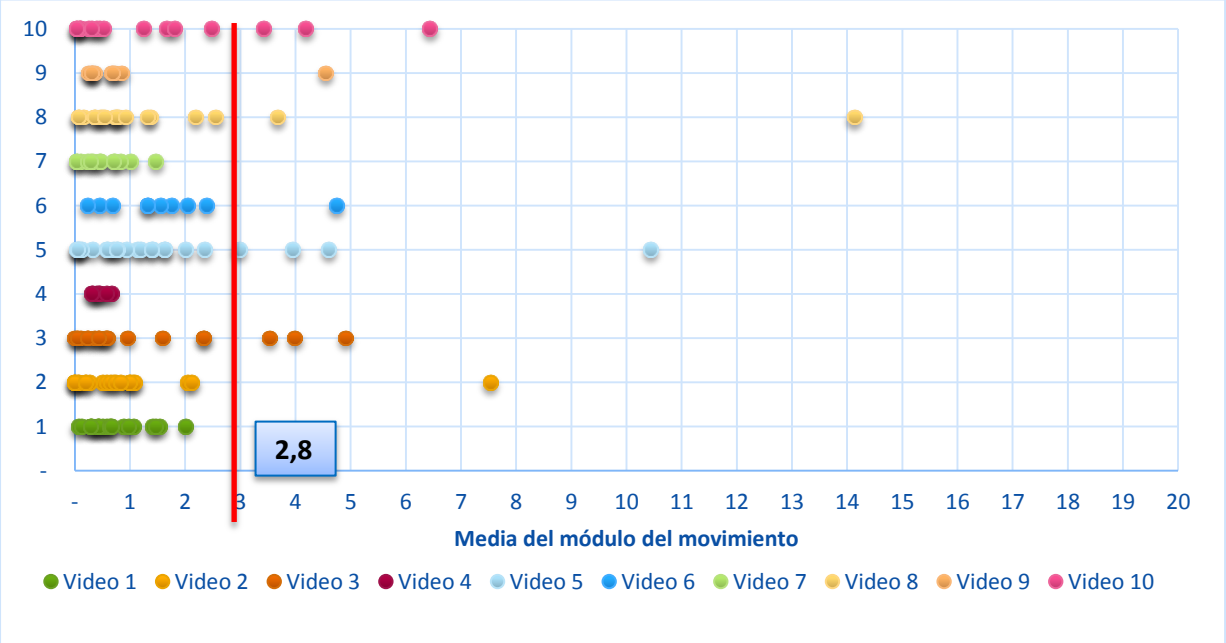

Figura 17. Detección de *shots* con th = 153 (60% de la imagen) y *long_mín* = 50 *frames* (2 seg).
Fuente: Propia

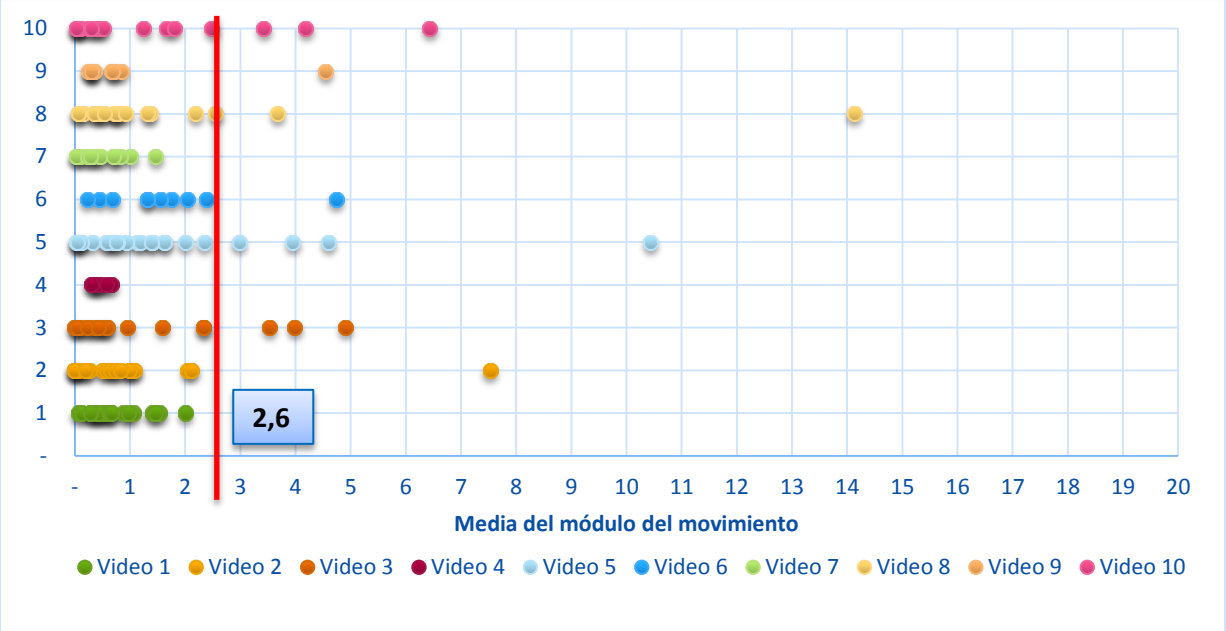

Figura 18. Detección de *shots* con th = 153 (60% de la imagen) y *long_mín* = 12 *frames* (≈ 0.5 seg). Fuente: Propia

El umbral resultado tras este análisis es el siguiente:

$$\boldsymbol{Umbral\ de\ movimiento} = \sum \boldsymbol{th_{mov}} = \mathbf{2.75}$$

El umbral de movimiento se considera una buena aproximación de forma generalista tras haber realizado un estudio sobre distintos casos, pero se tendrá que tener en cuenta que este umbral se verá influenciado por los siguientes aspectos:

1. Duración del *shot*
2. Rapidez del movimiento

Es decir, un movimiento rápido en un *shot* de corta duración va a tener más repercusión en el valor de la media de cantidad de movimiento que ese mismo movimiento en un *shot* de una duración mayor, ya que para este último se tendrá en cuenta, a la hora de hacer la media, el resto de movimiento del *shot*.

indexVideoShot

Cuando cada uno de los módulos descritos ha terminado de indexar al *shot* en cuestión con sus características concretas, se pasa al módulo final de indexación.

En este módulo final de indexado se recogen todas las etiquetas que se han asignado a dicho *shot* (y que estaban almacenadas en la clase *saveFeatures*) y se decide cuál será su representativa, escribiéndola en un archivo de texto a modo de registro. Este registro vendrá diferenciado, en función de los *shots*, con sus características etiquetadas.

La decisión sobre cuáles son las etiquetas finales o índices finales que deben atribuirse a cada *shot* se obtienen tras acceder al valor guardado por cada uno de los módulos de indexado por característica (*indexBGR*, *indexYCrCb*, *indexPeopleDetection*, *indexMovementDetection*) y guardando la etiqueta que en cada uno de ellos se había dado. Se hace uso de parámetros booleanos que se activan o desactivan en función de lo que recibe de estos módulos concretos.

Las etiquetas que se asignan tras este módulo son las recogidas por cada módulo de indexación específico. A modo de resumen, estas etiquetas serán:

BLUE	GREEN	RED
LUMINANCE	CHROMINANCE	
SINGLE PERSON	MULTIPLE PEOPLE	
CLOSEUP	BACKGROUND	
LOW MOVEMENT	HIGH MOVEMENT	

La combinación de estas etiquetas derivará en categorías concretas (establecidas específicamente para noticiarios, ya que es el género del *dataset* utilizado) que se definirán a partir de su estudio y análisis. Este punto se comenta en profundidad a continuación.

4.6 Categorías definidas para la indexación

4.6.1 Dataset

Para llevar a cabo las pruebas y la evaluación del algoritmo, cuyos resultados se verán en el siguiente capítulo, se hará uso de un conjunto de 10 vídeos de programas de noticiarios o informativos, pertenecientes a la cadena *Televisión Española*. Estos vídeos combinan diferentes contextos con variada información, por lo que de ella puede extraerse características distintas y realizar el análisis adecuado para ser capaces de indexarlos correctamente según su contenido visual.

Se pretende, mediante estas pruebas, obtener unos resultados aceptables que cumplan con el principal objetivo que se ideó al plantear este trabajo: obtener un **sistema de etiquetado automático**, a partir de características de bajo nivel, que sea capaz de discernir adecuadamente diferentes entornos que se plantean a priori, como son los reportajes de

exterior de los que son de interior, las escenas grabadas en estudio de las que están grabadas con luz natural, etc.

Dado que partíamos de un *dataset* en el que **no** se proporcionaban los *ground truths*, los cuales se creen imprescindibles para la evaluación del sistema, se generan estos mismos a partir de su visualización y experimentación. Con ellos, se realiza una primera ejecución completa del algoritmo, en la que se extraen las características y se etiqueta o dicho de otro modo, **indexa**, a nivel de *shot,* resultando registros con dichos índices para cada *shot* que el algoritmo ha diferenciado.

Se incluye una captura a modo de ejemplo del tipo de archivo de texto o registro que se genera tras la ejecución, el cual contiene los **índices por *shot* para cada vídeo** del conjunto de entrenamiento. A partir de estos registros o *logs*, un usuario será capaz de recuperar contenido multimedia mediante una *query*, como se ha comentado en el Capítulo 2 del estado del arte.

4.6.2 Creación de los *ground truths*

Se generan a partir de la visualización de los vídeos del conjunto utilizado durante el desarrollo, anotando etiquetas definitivas para cada *shot*.

Se debe hacer hincapié en que el objetivo final **no es anotar la característica observable en sí** (como serían el canal de color, canal de luminancia, etc.), **sino que se busca anotar la abstracción que se obtiene en un nivel superior para el conjunto de características observables.** A partir de esta abstracción se puede obtener un significado más tangible del contenido del vídeo.

Para ello, se va a crear un *ground truth* por cada uno de los vídeos dándole una etiqueta o índice a nivel de *shot* a modo de **categoría**, realizando la abstracción comentada para las características que se observan.

Para poder llevar esto a cabo con el algoritmo diseñado, se necesita el conocimiento previo de las distintas categorías que se quieren definir y según las cuales se indexará y guardará el vídeo, dejándolo indexado para una futura y más sencilla recuperación.

Por tanto, es necesario fijar unas reglas que establezcan una **relación directa entre características extraídas y etiqueta o índice final (categoría)**.

4.6.3 Creación de las reglas

Se razonan en este apartado las reglas seguidas para la creación de los *ground truths*, la indexación de los vídeos y la realización de sus pruebas.

El primer paso a llevar a cabo es definir las **categorías** que se diferenciarán en los **vídeos de noticiarios**, en cuyo estudio está centrado este trabajo:

Categorías		Descripción
Estudio	**Presentador**	Escena grabada en interior que, normalmente, da paso a la noticia en sí mediante un presentador que la anuncia e introduce.
	Entrevista / Debate	Escena grabada en interior donde interactúan dos o más individuos que discuten un tema de actualidad.
Noticia	**Reportero / Entrevistado**	Escena que puede ser grabada tanto en exterior como en interior que introduce un individuo, normalmente interactuando con el presentador del estudio, que detalla los hechos que se acontecen en la noticia que se presentará a continuación.
	Reportaje	Escena grabada tanto en exterior como en interior que muestra mediante imágenes de vídeo, ya sea en directo o pregrabadas, la noticia de la que se habla y se informa.

Tabla 1. Categorías definidas en los noticiarios y su descripción. Fuente: Propia

Estas categorías serán alcanzadas tras generar asociaciones en las características extraídas, en función de **tres pilares**: visualización y experimentación con los vídeos del *dataset*, así como del estudio de las tendencias seguidas en el estado del arte para este género (noticiarios) [8].
Las asociaciones establecidas entre etiquetas y categorías se definirán también en el capítulo a continuación.

4.7 Conclusiones

Como se ha podido observar a lo largo de este capítulo, se crea un sistema de indexado automático de vídeo a partir de sus características visuales globales. Este procedimiento se diseña e implementa por completo, desde su inicio a su fin, teniendo en cuenta ciertas

consideraciones que se han ido comentando y tomando las decisiones que se creen más oportunas para el fin que se persigue.

Todo este proceso se ejecuta sobre un conjunto de vídeos de noticiarios y, los resultados que ofrece en cuanto a su eficiencia y precisión se comentan en el capítulo que se muestra a continuación.

Capítulo 5. Evaluación del sistema

5.1 Introducción

En este capítulo se procede a realizar la evaluación del sistema propuesto, con el objetivo de medir su rendimiento y precisión.

Durante los últimos años, la evaluación de las técnicas enfocadas en la indexación y recuperación de contenido multimedia ha cobrado vital importancia[REF]. El crecimiento en este ámbito viene determinado por dos principales factores:

1. El incesante incremento en el número de sistemas de recuperación.
2. Un interés adicional en métodos de evaluación de dichos sistemas.

Estos dos factores vienen sesgados claramente por un punto en común: el desorbitante espacio de información que se encuentra en la red, lo que a veces es denominado como "*infoespacio*" [19], accesible desde múltiples focos y por infinidad de individuos. Además, y dada la tendencia exponencial que toma este camino, se hace rigurosamente necesario encontrar no sólo algoritmos que simplifiquen el proceso de indexación y recuperación, sino que estos sean lo más efectivos posible. En este punto nacen los **sistemas de evaluación en la recuperación de información**.

Existen para ello foros de evaluación anuales, donde uno de los más conocidos es el TREC (*Test REtrieval Conference*), en los que se introducen nuevas metodologías para la evaluación de dichos sistemas, promoviendo así la búsqueda de solución y mejora continua de esta problemática.

De forma general, las razones para evaluar la efectividad de estos sistemas giran en torno a las siguientes cuestiones:

- Identificar un sistema válido para el fin concreto que se persigue.
- Monitorizar y evaluar la eficacia del sistema.
- Evaluar el sistema para proponer mejoras.
- Obtener *inputs* que sirvan como método de análisis entre el riesgo-beneficio que ofrece un sistema de información.
- Determinar los efectos que conlleva realizar cambios sobre un sistema informacional ya existente y en funcionamiento.

También cabe destacar que existe una ligera diferencia entre aquellas medidas realizadas para evaluar sistemas enfocados a aspectos académicos de aquellos más ligados al uso comercial. Ambos buscan medir el rendimiento y eficacia del sistema, pero las evaluaciones sobre

sistemas destinados a uso comercial hacen gran hincapié en la disponibilidad y su nivel de confianza.

Además, para propósitos académicos pueden crearse sistemas de control que minimicen los errores en los datos. Sin embargo, para sistemas comerciales, no existe un control posible sobre los usuarios y se debe poner especial cuidado para asegurar que los datos recogidos son significativos.

5.2 Metodología de evaluación y métricas empleadas

Este apartado se presenta como una breve explicación previa al desarrollo de las pruebas con el sistema propuesto.

Esta explicación previa responde a la necesidad de asentar las bases iniciales de las que se parte, donde se definirán ciertos criterios que habrán de seguirse durante el tiempo de trabajo. Estos criterios actuarán, en mayor o menor medida, como condicionantes en las decisiones a tomar.

Es importante tener en cuenta que se ha desarrollado un **sistema de alto nivel, que surge como una primera aproximación** a un sistema de indexación automática de vídeo.

Por ello, el sistema a desarrollar será limitado y dispondrá de funcionalidades básicas, con vistas a que estos primeros pasos sirvan de base para posteriores extensiones o especializaciones en el mismo.

Por esta razón, las pruebas que se realizan también tienen limitaciones, que se contemplan como casos especiales y que se explican a continuación, de modo que en un futuro trabajo fueran de los primeros aspectos a mejorar.

5.2.1 Pruebas

El algoritmo propuesto ha sido diseñado en base a los conocimientos adquiridos tras el estudio del estado del arte, así como de las etiquetas particulares que podrían obtenerse para el tipo de vídeos que se planteaba como caso de análisis: los noticiarios. Tras dicho diseño y su posterior implementación, se procede a realizar las pruebas del algoritmo, con el fin de evaluar el rendimiento ofrecido.

Como se comentó en el capítulo anterior, para que el sistema establezca relaciones entre características extraídas y etiqueta que otorgar al shot en cuestión, es necesario establecer unas asociaciones previas, que son las siguientes:

Información de luminancia	Color dominante	Índice
Luminance	*Red*	**Interior**
	Green	**Exterior**
	Blue	**Exterior**
Chrominance	*Red*	**Interior**
	Green	***Exterior/Interior**
	Blue	***Exterior/Interior**

Entorno de grabación	Detección de personas	Detección de movimiento	Tipo de plano	Categoría asignada en base a la asociación
Exterior	*Single person*	*Low*	*Close up*	**Noticia: Reportero / Entrevistado**
			Background	**Noticia: Reportaje**
		High	*Close up*	**Noticia: Reportero / Entrevistado**
			Background	**Noticia: Reportaje**
	Multiple person	*Low*	*Close up*	**Noticia: Reportaje**
			Background	
		High	*Close up*	
			Background	
Interior	*Single person*	*Low*	*Close up*	**Estudio: Presentador *Noticia: Reportero / Entrevistado**
			Background	
		High	*Close up*	
			Background	
	Multiple person	*Low*	*Close up*	**Estudio: Entrevista / Debate *Noticia: Reportaje**
			Background	
		High	*Close up*	
			Background	

**Se comenta a continuación estos casos especiales*

Tabla 2. Tabla 5.2. Asociaciones preestablecidas entre características para asignar categoría. Fuente: Propia

Es importante tener presente que estas categorías que se han definido son muy generales y que podrían diferenciarse muchas otras más específicas. Sin embargo, para ello sería necesario un sistema mucho más potente en el período de la extracción de características, ya que de sus asociaciones depende en gran medida la identificación de una u otra categoría.

Aquí, por ejemplo, la categoría reconocida como "**Noticia: Reportaje**" es inmensamente amplia. En este tipo de categoría aparecen muchos tipos de escenarios, que pueden diferir completamente entre ellos aunque se refieran a un mismo tema (como ya se había introducido en el Capítulo 4, un reportaje se relaciona con un contenido *multishot*). Por ejemplo, un reportaje puede tratar una noticia sobre el cambio climático y mostrar en ella distintos paisajes de la naturaleza que representan este aspecto, lo que conllevaría escenas grabadas en exterior, donde no aparecerían personas y que, posiblemente, tendrían un alto movimiento en la imagen. Del mismo modo, podrían entremezclarse con dichas escenas, otras que recogieran imágenes de ruedas de prensa o congresos sobre dicho cambio climático, lo cual supondría escenas de interior, con muchas personas presentes que interactúan entre ellas.

Casos especiales

Se debe hacer mención a una serie de consideraciones en ciertos casos especiales que se han encontrado durante la creación de los *ground truths* y la generación de categorías en base a las características extraídas por el sistema propuesto.

Información de luminancia	Color dominante	Índice
Chrominance	*Red*	**Interior**
	Green	***Exterior/Interior**
	Blue	***Exterior/Interior**

Entorno de grabación	Detección de personas	Detección de movimiento	Tipo de plano	Categoría asignada en base a la asociación
Interior	*Single person*	*Low*	*Close up*	**Estudio: Presentador** ***Noticia: Reportero / Entrevistado**
			Background	
		High	*Close up*	
			Background	
	Multiple person	*Low*	*Close up*	**Estudio: Entrevista / Debate** ***Noticia: Reportaje**
			Background	
		High	*Close up*	
			Background	

Tabla 3. Casos especiales en categorías asignadas. Fuente: Propia

En primer lugar, el algoritmo va a encontrar dificultades a la hora de discernir adecuadamente entre escenas grabadas en interior y escenas grabadas en exterior cuando se trata de los siguientes casos:

1. ***Chrominance – Green***
2. ***Chrominance - Blue***

Ya que por sí solos no son capaces de concretar con exactitud si la escena corresponde con interior o exterior. Esto podría reforzarse en una siguiente etapa del trabajo mediante la extracción de más características que incidan en este aspecto.

Además existen ciertos casos en los que el sistema, dadas las características extraídas, no va a ser capaz de diferenciar correctamente entre estas dos categorías:

1. **Estudio: Presentador** y **Noticia: Reportero / Entrevistado.**
2. **Estudio: Entrevista / Debate** y **Noticia: Reportaje.**

La asociación de características que presentan las categorías que llevan a confusión es prácticamente idéntica dadas las características extraídas por nuestro sistema. Se presenta un ejemplo a continuación:

Figura 19. Ejemplo gráfico de un caso especial: Estudio: Presentador y Noticia: Reportero / Entrevistado. Fuente: Propia

Como puede observarse en la figura anterior y que se corresponde con el primer caso mencionado, ambas imágenes contienen una escena grabada en interior, con una sola persona presente en ella, que corresponde con lo considerado un primer plano y donde el movimiento es bajo. Sin embargo, la primera hace referencia a un plano de estudio con presentador y la segunda a una toma del reportaje durante la noticia.
Por tanto, para ser capaz de diferenciar entre dichas categorías en este tipo de casos concretos, será necesario hacer uso de un mayor conjunto de características extraídas en la imagen.

Del mismo modo ocurre con el **segundo caso** mencionado, donde ambas categorías se asignan en base a una posible misma asociación de características de las que extrae el sistema desarrollado.

En conclusión, en estos casos especiales comentados anteriormente, el sistema realizará **excepciones** a la hora de hacer el *ground truth* y **admitirá para el posterior *matching* esta posible confusión**, planteando como trabajo futuro una mayor profundidad de estudio en este sentido.

Hay que tener en cuenta que este tipo de excepciones deben admitirse en un sistema cuya finalidad es realizar una primera aproximación a un algoritmo de indexación automática como el que aquí se plantea y cuyo estudio es una línea de investigación hoy en día dada su amplitud y dificultad.

5.2.2 Resultados

El algoritmo se ha diseñado, como se explicaba en el Capítulo 4, de tal modo que procese distintos vídeos de manera consecutiva y vaya generando los registros correspondientes con la asignación de índices o etiquetas. Por tanto, es posible llevar a cabo un análisis de resultados combinando las siguientes fuentes:

- **Registros o *logs*** generados por el algoritmo con las etiquetas correspondientes.
- ***Ground truths*** generados tras la visualización del conjunto de vídeos utilizados durante el desarrollo del algoritmo.
- **Categorías** definidas a priori en los vídeos de noticiarios y cuya abstracción da un sentido al contenido de los mismos.

Para llevar a cabo la evaluación de resultados, se diseñará una **plantilla de evaluación.** Esta plantilla constará de los siguientes campos:

1. ***Shot***: número de *shots* en los que el algoritmo divide al vídeo.
2. ***Logs***: valores de índices o etiquetas que el algoritmo creado genera tras su ejecución.
3. **Categoría**: a partir de las reglas creadas, valor que se obtiene al combinar las características reflejadas en los *logs* según dichas reglas.
4. ***Ground truth***: valor para la categoría que se ha observado.
5. ***Matching***: comprobación de si ambos resultados concuerdan (entre categoría a partir de lo etiquetado en los *logs* y la categoría definida en el *ground truth*).
6. ***Score***: Puntuación que se le otorga al funcionamiento del algoritmo para el *shot* concreto que se analiza, resultando en correspondencia correcta (TP – *True Positive*) o correspondencia incorrecta (FP – *False Positive*).

Shot	*Logs*	*Category*	*Ground truth*	*Matching*	*Score*
1	Shot dominant color: Blue Brightness and color information in the shot: Chrominance Single person in the shot Type of shot: Close up Movement in the video shot: Low	Estudio: Presentador	Estudio: Presentador	SÍ	TP
2	Shot dominant color: Blue Brightness and color information in the shot: Chrominance Multiple people in the shot Type of shot: Close up Movement in the video shot: Low	Estudio: Presentador	Estudio: Presentador	SÍ	TP
3	Shot dominant color: Green Brightness and color information in the shot: Chrominance Single person in the shot Type of shot: Close up Movement in the video shot: Low	Noticia: Reportero /Entrevistado	Noticia: Reportero /Entrevistado	SÍ	TP
4	Shot dominant color: Green Brightness and color information in the shot: Chrominance Single person in the shot Type of shot: Close up Movement in the video shot: Low	Noticia: Reportero /Entrevistado	Noticia: Reportaje	NO	FP
5	Shot dominant color: Green Brightness and color information in the shot: Chrominance Multiple people in the shot Type of shot: Close up Movement in the video shot: Low	Noticia: Reportaje	Noticia: Reportaje	SÍ	TP

Tabla 4. **Ejemplo de la plantilla de evaluación con la información sobre uno de los vídeos del *dataset* para sus primeros *shots*. Fuente: Propia**

Scores

En función de las distintas combinaciones posibles entre los resultados del *ground truth* y los registros que crea el algoritmo tras su ejecución se realizarán las siguientes clasificaciones:

- ***True Positive***: Se detecta TP cuando el sistema indexa con una categoría concreta y acierta.

- ***False Positive***: Se detecta FP cuando el sistema indexa con una categoría concreta y se equivoca.

TP y FP son medidas absolutas, que suelen ir acompañadas de otras medidas relativas (True Negative o TN, False Negative o FN). Sin embargo, en este problema que se plantea la decisión a tomar es **no binaria** y por ello no se pueden deducir de manera clara estas medidas relativas. Por tanto, para ser capaces de evaluar el rendimiento del sistema de una manera eficaz, se van a definir dos métodos:

Precision

$$Precision = \frac{TP}{TP + FP}$$

El valor de *Precision* aumenta cuando hay pocos falsos positivos. Mide que las instancias clasificadas como categoría "*x*" sean realmente de la categoría "*x*", aunque haya instancias de dicha clase que se clasifiquen como otra diferente. En ella intervienen las dos medidas absolutas comentadas anteriormente: TP y FP.
A continuación se muestra el valor de *Precision* obtenido para el conjunto de *dataset* utilizados:

	Total Shots	*TP*	*FP*
Vídeo 1	23	9	14
Vídeo 2	27	13	14
Vídeo 3	21	12	9
Vídeo 4	9	8	1
Vídeo 5	29	12	17
Vídeo 6	10	10	0
Vídeo 7	15	11	4
Vídeo 8	27	8	19
Vídeo 9	11	6	5
Vídeo 10	19	15	4
TOTAL	191	104	87

Tabla 5. Valores de los scores TP y FP recogidos. Fuente: Propia

$$\boldsymbol{Precision} = \frac{TP_{total}}{TP_{total} + FP_{total}} = \frac{104}{104 + 87} = \mathbf{54,45\ \%}$$

Este valor obtenido como resultado ha sido dado **al analizar cada uno de los *shots* de los vídeos de manera independiente**, teniendo todos la misma ponderación y ejerciendo el mismo peso en el resultado global.

Sin embargo, podría analizarse un **caso particular desde otra perspectiva**. Como ya se ha comentado, la categoría asignada como **Noticia: Reportaje** puede aunar una gran variedad de *shots* diferentes para un mismo reportaje (el cual corresponde dentro de la taxonomía a un segmento *multishot*), por lo que dentro de un mismo reportaje pueden calificarse *shots* de muy variadas características. Por tanto, para un mismo reportaje están incluidos una gran cantidad de *shots*, que pueden poseer características muy dispares aunque se refieran a un mismo tema.

De este modo, al analizar a nivel de *shot*, si uno de los vídeos contiene un reportaje de una duración superior a la media y con alguna característica que resulte problemática (oclusiones, sombras, mala iluminación, movimientos bruscos de la cámara, etc.) y que se repita durante el desarrollo del mismo, ésta va a estar penalizando continuamente al vídeo bajo estudio. Lo mismo ocurrirá si algún vídeo posee características muy claras y distintivas que el sistema sea capaz de etiquetar correctamente, lo que contribuirá a sobrepuntuar.

Dado que, además, la base de datos de vídeos utilizados no era muy grande (hecho en gran parte condicionado por el requerimiento de tener que construir los *ground truths* manualmente para la evaluación), se cree que una **manera alternativa** de evaluar el rendimiento del algoritmo teniendo en cuenta estos aspectos es **ponderando a cada uno de los vídeos por el peso que tiene en la evaluación total**. Es decir, el número de *shots* que posee cada vídeo es independiente para el resultado global, que contempla al valor del vídeo como una unidad indivisible.

Se propone este método de evaluación como una **medida basada en el *precision* clásico, pero que aplica y tiene en cuenta esta ponderación adicional** que viene dada por la forma de evaluar los vídeos del *dataset* (a nivel de *shots*). Pueden encontrarse varias referencias en la bibliografía sobre casos de algoritmos que proponen sus propios métodos de evaluación, ajustados en parte a las condiciones del sistema que proponen.

Para ello, los valores de la tabla anterior, es decir los TP' o FP' correspondientes a cada vídeo, serán ponderados por el número de *shots* presentes en el vídeo en concreto. Tras esta modificación de perspectiva, los resultados son los siguientes:

	Total Shots	TP'	FP'
Vídeo 1	23	9/23 = 0,39	14/23 = 0,61
Vídeo 2	27	13/27 = 0,48	14/27 = 0,52
Vídeo 3	21	12/21 = 0,57	9/21 = 0,43
Vídeo 4	9	8/9 = 0,89	1/9 = 0,11
Vídeo 5	29	12/29 = 0,41	17/29 = 0,59
Vídeo 6	10	10/10 = 1,00	0/10 = 0,00
Vídeo 7	15	11/15 = 0,73	4/15 = 0,27
Vídeo 8	27	8/27 = 0,30	19/27 = 0,70
Vídeo 9	11	6/11 = 0,55	5/11 = 0,45
Vídeo 10	19	15/19 = 0,79	4/19 = 0,21
TOTAL	191	6,1	3,9

Tabla 6. Valores de los scores TP' y FP' recogidos, bajo el método de evaluación propuesto basado en ponderación por *shots*. Fuente: Propia

$$\boldsymbol{Weighted\ precision} = \frac{TP_{total}}{TP_{total}+FP_{total}} = \frac{6{,}1}{6{,}1+3{,}9} = \mathbf{61{,}12\ \%}$$

Como se puede apreciar, la **ponderación** que se otorga con este método de evaluación **da lugar a una mejora de los resultados**, pero sin ser una mejora excesiva ya que se sigue teniendo en cuenta tanto los casos que penalizan como los que sobrepuntúan.

Matriz de confusión

Se utiliza como una herramienta útil para la visualización del desempeño de un algoritmo, donde cada columna de la matriz representa el número de predicciones o estimaciones de cada clase y cada fila representa a las instancias en la clase real. Es decir, en el caso que se plantea, las columnas se corresponderán con las distintas categorías que el sistema devuelve tras su ejecución mientras que, cada fila reflejará las categorías reales para los noticiarios que se han definido y anotado en el *ground truth* en este trabajo.

El rendimiento que se obtiene con el algoritmo viene dado por la diagonal de la matriz. En dicha diagonal deben acumularse los valores más altos, que supondrá que ha habido un número mayor de aciertos en las correspondencias entre categoría real y categoría predicha.

Siendo las categorías reales:

- **Categoría 1**: Estudio - Presentador
- **Categoría 2**: Estudio - Entrevista/Debate
- **Categoría 3**: Noticia - Reportero/Entrevistado
- **Categoría 4**: Noticia – Reportaje

Y las categorías predichas o estimadas:

- **Categoría 1'**: Estudio - Presentador
- **Categoría 2'**: Estudio - Entrevista/Debate
- **Categoría 3'**: Noticia - Reportero/Entrevistado
- **Categoría 4'**: Noticia – Reportaje

En primer lugar se muestra una tabla a modo de resumen que contiene los datos según el número de *shots* por categoría y por vídeo, resaltados en el mismo color aquellos que pertenecen a las mismas categorías:

	Categoría 1	Categoría 2	Categoría 3	Categoría 4	*Total Shots* por vídeo
Vídeo 1	3	2	4	14	23
Vídeo 2	0	0	3	24	27
Vídeo 3	0	2	2	17	21
Vídeo 4	2	0	6	1	9
Vídeo 5	0	0	4	25	29
Vídeo 6	1	1	0	8	10
Vídeo 7	0	0	1	14	15
Vídeo 8	1	0	5	21	27
Vídeo 9	1	0	4	6	11
Vídeo 10	2	1	2	14	19
Total *shots* *por categoría*	10	6	31	144	191

Tabla 7. Número de shots en función de categoría y vídeo (*ground truth*). Fuente: Propia

Y, a continuación, se muestra la matriz de confusión propiamente dicha, donde pueden verse estas correspondencias

		Categorías estimadas o predichas			
		Categoría 1'	Categoría 2'	Categoría 3'	Categoría 4'
Categorías reales	Categoría 1 (10)	6	1	3	0
	Categoría 2 (6)	0	3	0	3
	Categoría 3 (31)	1	0	28	2
	Categoría 4 (144)	0	2	75	67

Tabla 8. Matriz de confusión. Fuente: Propia

Como puede observarse en la tabla anterior, la matriz muestra que el algoritmo concentra, en conjunto, la mayor parte de los aciertos en su diagonal, vistos como correspondencias correctas entre categorías definidas o reales y categorías estimadas o predichas. Esto quiere decir que el sistema de indexación automática es capaz de reconocer de manera adecuada las características visuales globales extraídas y que, las reglas creadas mediante la asociación de las mismas tienen sentido y concuerdan con estas características.

Entrando más en detalle en los resultados que devuelve la matriz, se observa que la mayor parte de los *shots* se concentran en las categorías 3 y 4, **Noticia: Reportero/Entrevistado** y **Noticia: Reportaje**, respectivamente, siendo esta última la más recurrente. Esto es lógico, ya que la mayor parte del contenido dentro de un programa informativo gira en torno a los reportajes, donde ambas categorías entran en juego.

También se puede observar que el sistema acumula más fallos cuando tiene que intentar indexar correctamente esta categoría (**Noticia: Reportaje**) y clasificarla como tal. Concretamente, confunde en gran medida esta categoría con la de **Noticia: Reportero/Entrevistado** ya que, en gran parte de los reportajes, se dan muchas características similares a las de esta otra categoría y el sistema no siempre es capaz de diferenciarlas con la información de la que dispone. Esto se debe a que la variedad en reportajes es muy alta y, por tanto, sólo con un número mayor de características extraídas y su adecuada combinación será capaz de hacer un análisis más fino que ayude a mejorar el rendimiento en este aspecto.

Justificación de los resultados

En vista de los resultados presentados, y teniendo en cuenta las decisiones tomadas a la hora de afrontar ciertos inconvenientes mencionados en los casos especiales, se comentan a continuación dónde residen las principales carencias o debilidades del sistema propuesto.

En primer lugar y, tal y como se ha ido comentando, se ha creado un sistema de indexación automática de vídeo que es capaz de lo siguiente:

- **División correcta del vídeo en *shots* útiles.**

El algoritmo es capaz de seccionar un vídeo en sus distintos shots siguiendo una pauta bastante clara, de modo que se tengan diferenciados correctamente los reportajes de las escenas de estudio. Además, **la inclusión de efectos en la imagen, como cabeceras, subtítulos, etc.,** no perjudica el funcionamiento del sistema, ya que no se considera un

cambio suficientemente grande como para detectar nuevo *shot*, ni de dichos efectos se extraen características que perjudiquen a la asociación establecida para la creación de las reglas.

Figura 20. División útil de *shots* de un mismo vídeo; por orden descendiente: Estudio: Presentador, Noticia: Reportaje, Noticia: Reportero/Entrevistado. Fuente: Propia

Figura 21. Imágenes pertenecientes al mismo *shot* a pesar de la inclusión de efectos. Fuente: Propia

- **Capacidad de indexación automática en diferentes categorías genéricas.**

El sistema es capaz de extraer características y, mediante su combinación y con la ayuda inicial que se genera al crear las reglas y definir las categorías, pueda distinguir y diferenciar aspectos intrínsecos en la imagen para clasificarlas y almacenarlas de forma que quede representado y

etiquetado por ciertos índices que faciliten su posterior recuperación.

Estos aspectos constituyen claramente la **fortaleza** del sistema, el cual podría irse reforzando y ampliando en rigurosidad y eficiencia partiendo del algoritmo propuesto.

Por otro lado, las **debilidades** principales del algoritmo residen en ciertas situaciones; algunas de ellas se comentan a continuación:

- **Detección de personas ineficiente en ciertos casos,** requiriendo más funciones que refuercen su correcto desempeño, además de la integración con más características que la complementen para que la detección pueda resultar más efectiva**:**

Figura 22. Ejemplos de debilidades del algoritmo (I). Fuente: Propia

En las dos imágenes superiores, debido a las características que se observan durante dichos *shots*, en los que las personas aparecen con cascos/gorras y gafas de sol (oclusión parcial de las caras), las funciones de detección de caras que se implementaron partiendo de ciertas plantillas dadas por OpenCV (comentadas en el módulo *peopleDetection* del Capítulo 4) no son precisas en estos casos, ya que no detectan persona cuando sí debería o se no detecta a todos los individuos.
Mientras que esta falta de detección sucede en ocasiones como la comentada, también se da el caso de las imágenes inferiores de la figura, donde se puede observar que no aparece ninguna persona pero el algoritmo confunde ciertas características en ellas y detecta persona donde no la hay.

Este tipo de confusiones en la extracción de características (en este caso concreto, en la característica que recoge la detección de personas) hace que el conjunto de características extraídas siga la asociación establecida para una regla que no es correcta, pudiendo calificar un *shot* con una categoría que, en ocasiones, no es la que le corresponde.

También es importante mencionar que, dado el *dataset* del que se partía y dado que son pocas las ocasiones en las que en la imagen no aparecen personas, no se ha contemplado esta casuística durante la implementación del algoritmo, por lo que tampoco se ha tenido en cuenta a la hora de crear las reglas. Sin embargo, este podría ser un caso interesante para el trabajo futuro ya que muchos de los falsos positivos vienen dados por esta razón.

- **Necesidad de un mayor número de características en la extracción.**

Aunque no siempre esto es sinónimo de un mejor rendimiento en el sistema, es cierto que la inclusión de más características analizadas de un modo riguroso y combinadas adecuadamente aumenta la eficacia del sistema.

Por ejemplo, en las siguientes imágenes vemos dos claros casos en los que el *shot* debería ser calificado como **Noticia: Reportero/Entrevistado** pero, dado que en la escena aparece más de una persona y el sistema lo detecta así, la asociación establecida para las características extraídas es insuficiente en esta situación, calificando **Noticia: Reportaje** (si bien es cierto que forma parte del reportaje, no se da la situación que se pretende diferenciar).

Figura 23. Ejemplos de debilidades del algoritmo (II). Fuente: Propia

5.3 Conclusiones

A modo de conclusión final, en este capítulo se han definido los métodos de evaluación que han sido utilizados para realizar las pruebas del algoritmo creado de indexación automática de vídeo.

Además, se han descrito las distintas posibilidades de evaluación y se han justificado los métodos llevados a cabo, incidiendo en la razón de las decisiones que se han ido tomando a medida que ha avanzado el desarrollo del sistema.

Por tanto, se puede concluir que se ha **diseñado, desarrollado e implementado un sistema base de indexación automática de vídeo, genérico y sencillo, pero que recoge el proceso completo necesario para la indexación por medio de la extracción de características globales visuales**. Del mismo modo, se han explicado las carencias que sufre y en el capítulo que sigue a continuación se ahondará en cómo éstas pueden ser solventadas de cara a un trabajo futuro.

Capítulo 6. Conclusiones y trabajo futuro

6.1 Conclusiones

Este proyecto se basaba en la creación de un **sistema de indexación automática de vídeo**, que fuera capaz de extraer **características visuales globales** sobre un caso de estudio planteado con vídeos de **noticiarios**. A partir de esas características extraídas, el sistema atribuirá índices al contenido multimedia, de forma que quede etiquetado para una futura y más sencilla recuperación (*retrieval*).

Para este fin, se llevó a cabo un estudio sobre la bibliografía presente en el estado del arte sobre indexación de vídeo, sus principales métodos y técnicas implementados hasta día de hoy, así como las tendencias y retos a los que debe hacer frente.

Tras analizar en profundidad dicho estado del arte, se propuso el diseño de un **algoritmo a alto nivel**, surgiendo como una **primera aproximación** a un sistema de indexación automática de vídeo, con una estructura limitada y funcionalidades básicas, pero que sirven de **base para posteriores extensiones o especializaciones** en el mismo.

A partir de los resultados obtenidos en las pruebas y su posterior evaluación, se han podido establecer una serie de **conclusiones principales**:

- Para cualquier sistema basado en el procesado de vídeo es totalmente **imprescindible ser meticuloso en las etapas iniciales**, haciendo hincapié en la segmentación, en este caso temporal, para **obtener unidades homogéneas** sobre las que ir desarrollando los siguientes pasos. Esto es así ya que la bondad de todo el sistema que se realice a continuación dependerá en gran medida de este aspecto.

- La adecuada **extracción de características** en la imagen así como el **correcto análisis** que de ellas se derive es fundamental para poder indexar el contenido multimedia de una forma útil y que, por tanto, agilice el proceso de recuperación de vídeo, a cuyo fin se destina esta indexación automática.

- El enfoque proporcionado por el algoritmo implementado, basado en la indexación particular de vídeos de noticiarios, ha permitido poder definir un perímetro de actuación en cuanto a **categorías en función del contenido**. Esto se ha hecho de este modo dado que el sistema era una primera aproximación y, por tanto, era limitado, pero ha servido para asentar las bases iniciales que derivarían en un algoritmo robusto de etiquetado automático de vídeo.

En vista de lo expuesto en estas líneas y a modo de conclusión global:

Se ha creado un sistema de indexación automática de vídeo a muy alto nivel, con funcionalidades básicas, centrado en la extracción de características visuales globales en la imagen y la combinación de las mismas, de modo que pueda obtenerse una abstracción en el contenido y pueda ser identificado con escenarios reales. En este caso, se tomado como estudio vídeos de noticiarios y se han realizado pruebas sobre ellos para evaluar el rendimiento del sistema, el cual sirve como primera aproximación y debe mejorarse para funcionar de modo más robusto y general.

6.2 Trabajo futuro

Las líneas de trabajo futuro principales en las que se considera que se podría profundizar más a partir de la idea expuesta son:

- **Eficiencia:**

Se debería conseguir resultados más eficientes que los obtenidos, aplicando para ello mejores y más complejas técnicas de extracción de características que, con un análisis adecuado y la combinación de las mismas, harían mejorar el sistema al completo, siendo más riguroso en la indexación y registrando el contenido multimedia de una forma más práctica y ordenada.

- **Enfoques y aplicaciones:**

Dada la duración finita de este proyecto, no se ha podido profundizar más en la búsqueda de otros enfoques, pero como trabajo futuro se tendría muy en cuenta introducir nuevos módulos y funcionalidades en el algoritmo implementado, sobre todo centrados en la extracción de nuevas características.

Una de estas características sería la textura en la imagen, la cual proporciona información que, analizada de forma rigurosa, puede proporcionar resultados aplicables en muchos escenarios. De hecho, una primera aproximación de este módulo se ha empezado a implementar y gira en torno al cálculo del gradiente y su información asociada, como tipos de bordes, regulares e irregulares, de los objetos en una imagen y orientación de los mismos (bordes regulares suelen tener direcciones marcadas en ciertos ángulos, como en 0° y 90°, mientras que los bordes irregulares tienen direcciones más dispares o desiguales). Para ello se utilizan implementaciones basadas en filtros *Gabor*.

En última instancia, todo este proceso de extracción e indexado de características podría contemplar ser trasladado a una implementación en tiempo real.

Bibliografía

[1] Muneesawang, P., Zhang, N., & Guan (2014). L. Multimedia Database Retrieval, Springer International Publishing.

[2] Ansari, A., & Mohammed, M. H. (2015). Content based Video Retrieval Systems-Methods, Techniques, Trends and Challenges. *International Journal of Computer Applications*, *112*(7).

[3] Asghar, Muhammad Nabeel, Fiaz Hussain, and Rob Manton. "Video indexing: a survey." *framework* 3.01 (2014).

[4] Gitte, M., Bawaskar, H., Sethi, S., & Shinde, A. (2014). Content based video retrieval system. *International Journal of Research in Engineering and Technology*, *3*(06).

[5] Snoek, C. G., & Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia tools and applications*, *25*(1), 5-35.

[6] Del Fabro, M., & Böszörmenyi, L. (2013). State-of-the-art and future challenges in video scene detection: a survey. *Multimedia systems*, *19*(5), 427-454.

[7] Hu, W., Xie, N., Li, L., Zeng, X., & Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *41*(6), 797-819.

[8] Valdés, V., & Martínez, J. M. (2012). On-line video abstract generation of multimedia news. *Multimedia Tools and Applications*, *59*(3), 795-832.

[9] Mitisha Narottambhai Patel and Purvi Tandel. Article: A Survey on Feature Extraction Techniques for Shape based Object Recognition. *International Journal of Computer Applications* 137(6):16-20, March 2016. Published by Foundation of Computer Science (FCS), NY, USA.

[10] Otávio A. B. Penatti, Eduardo Valle, and Ricardo da S. Torres. 2012. Comparative study of global color and texture descriptors for web image retrieval. *J. Vis. Comun. Image Represent.* 23, 2 (February 2012), 359-380.

[11] Patel, B. V., & Meshram, B. B. (2012). Content based video retrieval systems. *arXiv preprint arXiv:1205.1641*.

[12] Manjunath, B. S., Ohm, J. R., Vasudevan, V. V., & Yamada, A. (2001). Color and texture descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, *11*(6), 703-715.

[13] C. Harris and M. Stephens (1988). "A combined corner and edge detector". *Proceedings of the 4th Alvey Vision Conference*. pp. 147–151.

[14] Canny, J., *A Computational Approach To Edge Detection*, IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679–698, 1986.

[15] Alonso, O., Strötgen, J., Baeza-Yates, R. A., & Gertz, M. (2011). Temporal Information Retrieval: Challenges and Opportunities. *TWAW*, *11*, 1-8.

[16] Lucas, B. D., & Kanade, T. (1981, August). An iterative image registration technique with an application to stereo vision. In *IJCAI* (Vol. 81, pp. 674-679).

[17] Zhang, T., & Kuo, C. J. (2013). *Content-based audio classification and retrieval for audiovisual data parsing* (Vol. 606). Springer Science & Business Media, 9-11.

[18] Klaser, A., Marszałek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference* (pp. 275-1). British Machine Vision Association.].

[19] Kowalski, G. (2010). *Information retrieval architecture and algorithms*. Springer Science & Business Media.

[20] Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Image analysis* (pp. 363-370). Springer Berlin Heidelberg.

[21] Yang, K., Cai, Z., & Zhao, L. (2013). Algorithm Research on Moving Object Detection of Surveillance Video Sequence. *Optics and Photonics Journal*, *3*(02), 308.