ARTÍCULOS

Cuadernos de Documentación Multimedia
ISSN: 1575-9733

http://dx.doi.org/10.5209/CDMU.62805

EDICIONES COMPLUTENSE

Segmentación de vídeos informativos en televisión : de la práctica profesional a la identificación automática

Jorge Caldera-Serrano[1]; Carmen Caro-Castro[2]

Recibido: 1 de septiembre de 2018 / Aceptado: 27 de diciembre de 2018

Resumen: Se analizan las diferentes maneras en el que el documentalista realiza la segmentación del vídeo para así identificar la unidad discursiva mínima de análisis en los departamentos de documentación de las cadenas televisiva. De forma paralela, y atendiendo al cambio de paradigma digital en la producción en televisión, se analizan las principales fórmulas de segmentación automática de vídeo desarrolladas por medio de la inteligencia artificial. Una vez determinada ambas realidades se intenta implementar el mejor método de segmentación entre lo humano y lo automático, para que tenga su utilidad en los sistemas de información documental de las televisiones. Para ello se ofrecen las principales líneas de trabajo en segmentación semántica de vídeo.
Palabras clave: Segmentación de vídeo / Microdiscurso audiovisual / Análisis de contenido / Televisión / segmentación semántica

[en] Information video segmentation on television: the profesional practice to the automatic indentification

Abstract: It analyzes the different ways in which documentary video segmentation performed in order to identify the minimal discourse unit of analysis in documentation departments of television channels. In parallel, and in response to the digital paradigm shift in television production, analyzes the main formulas developed automatic video segmentation through artificial intelligence. Having determined both realities are trying to implement the best method of segmentation between human and machine, to have its usefulness in information system television documentary. This will provide the main lines of work in semantic video segmentation.
Keywords: Video segmentation / audio-visual mike-speech / content analysis / televisión / semantic segmentation

Cómo citar: Caldera-Serrano, Jorge; Caro-Castro, Carmen (2019). Segmentación de vídeos informativos en televisión: de la práctica profesional a la identificación automática. *Cuadernos de Documentación Multimedia*, 30, 1-17.

[1] Departamento de Información y Comunicación. Universidad de Extremadura (España)
E-mail: jcalser@unex.es
[2] Departamento de Biblioteconomía y Documentación. Universidad de Salamanca (España)
E-mail: ccaro@usal.es

1. Introducción

La producción digital ha modificado los circuitos de producción en las cadenas de televisión. El cambio ha hecho posible que los periodistas puedan acceder a los contenidos que los servicios de documentación han almacenado en dispositivos digitales. De esta forma, ellos mismos pueden buscar la información y organizarla, además de montar, locutar y postproducir las noticias (López de Quintana, 2007). El nuevo formato, junto con el uso cada vez más generalizado de Internet, ha supuesto además que no solo los periodistas accedan a los contenidos audiovisuales custodiados por las empresas. La generalización del acceso "a la carta" en la Web y el interés por los sistemas de recomendación de programas hacen cada vez más necesario el acceso a la información audiovisual en formatos manejables para los usuarios finales. Formatos que permitan la exploración y el acceso tanto a programas completos como, de forma no secuencial, a fragmentos con contenidos de su interés.

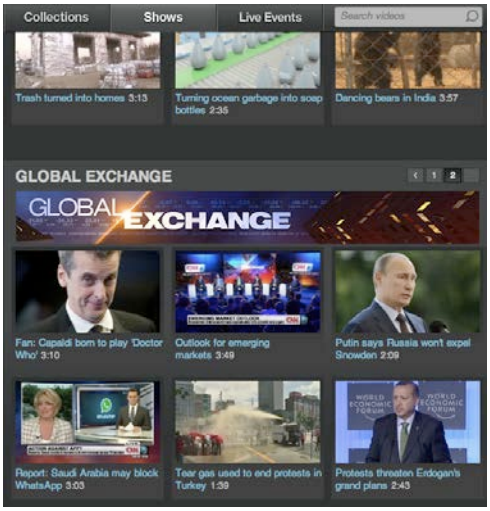

Figura 1. Ejemplo URL de acceso a contenidos televisivos
http://edition.cnn.com/video/

Una labor imprescindible para poder facilitar este acceso es el análisis de contenido de los documentos audiovisuales. Esta compleja tarea ha sido tradicionalmente realizada en las empresas televisivas por profesionales encargados de asignar metadatos a las unidades documentales – programas, noticias o secuencias – que se consideraban potencialmente útiles para una posterior utilización. El cambio generalizado a los soportes digitales y el incremento en la eficacia del procesamiento automático han posibilitado la utilización de herramientas innovadoras que permiten la segmentación del vídeo, su indización (selección de fotogramas representativos o keyframes) y su presentación al usuario para facilitar la exploración y la recuperación eficiente de datos visuales.

En la última década, el espectacular incremento de la creación, difusión y consumo de contenidos audiovisuales ha estimulado el desarrollo de técnicas de análisis automático de vídeo. A la producción de las, cada vez más numerosas, cadenas de televisión se suman la multitud de contenidos multimedia creados por particulares e instituciones y difundidos por Internet. Al ampliarse el rango de aplicación, por la necesidad de describir y organizar esta información para hacerla accesible a los usuarios de forma rápida y sencilla, ha crecido también el interés por la investigación en este campo. Hu, et al. (2011) señalan dos hitos de especial relevancia:

1).- Desde 2001 el National Institute of Standards and Technology han financiado la conferencia anual Text Retrieval Conference (TREC) Video Retrieval Evaluation (TRECVid), que en 2003 se independizó de TREC. Su objetivo primordial es promover el progreso en el análisis de contenido y la recuperación de información de vídeos digitales mediante una valoración de los sistemas basada en métricas abiertas. Las evaluaciones relacionadas en TRECVID son "estilo laboratorio", intentando modelar situaciones reales o componentes significativos de las tareas implicadas en dichas situaciones.

Hasta 2003, TRECVID utilizaba datos de una pequeña colección de fuentes profesionales – cadenas de informativos, productores de programas de TV y sistemas de vigilancia – que imponían límites relativos al estilo de los programas, contenido, calidades de producción, idioma, etc. A partir de entonces sus contenidos se han ido diversificando: entre 2003 y 2006 se utilizaron emisiones de noticias (en inglés, árabe y chino) y también vídeo no editado proporcionado por la BBC. Entre 2007 y 2009 se analizaron programas educativos, culturales, magacines de noticias y documentales facilitados por el Netherlands Institute for Sound and Vision. En 2010 TRECVID incorporó un nuevo conjunto de vídeos caracterizados por un alto grado de diversidad en cuanto a autor, contenido, estilo, calidades de producción, medio de grabación/codificación original, idioma, etc., como es habitual en muchos sitios web que distribuyen vídeo. Las tareas que se han realizado durante estos años con este material incluyen la segmentación, búsqueda, extracción de características visuales, detección de copias, identificación de ejemplares conocidos e indización semántica.

2).- El desarrollo de estándares cuya finalidad es asegurar la compatibilidad entre sistemas de descripción de contenidos de vídeo, con el objetivo de facilitar el desarrollo de algoritmos de búsqueda más rápidos y precisos. Los estándares más importantes son los promovidos por Moving Experts Group (MPEG)[3] y el TV-Anytime[4]. Dentro de la familia MPEG, en 2002 se publicó oficialmente el MPEG-7, utilizado en numerosas aplicaciones e investigaciones para extractar características con objeto de clasificar contenidos o para describir objetos de vídeo.

[3] http://mpeg.chiariglione.org/
[4] http://www.etsi.org/technologies-clusters/technologies/broadcast/tv-anytime

MPEG-7 establece una representación estándar de la información audiovisual que permite la descripción de contenidos por componentes estructurales (formas, colores, texturas, movimientos, sonidos) y por palabras clave (quién, qué, cuándo, dónde). Esta información se asocia de forma natural a los contenidos audiovisuales comprimidos por los codificadores MPEG-1 (almacena y descarga archivos audiovisuales), MPEG-2 (televisión digital) y MPEG-4 (codifica audio y vídeo en forma de objetos), pero se ha diseñado para que sea independiente del formato del contenido.

Este trabajo desea facilitar un aporte sobre los principales métodos de segmentación de vídeo. Se analizarán las diferentes unidades documentales y microdiscursos que han de ser analizados en los sistemas digitales de información y cómo el documentalista lleva a cabo la fragmentación de vídeo. En definitiva, definir cuáles son los elementos importantes que se deben describir y cómo agrupar imágenes para una descripción documental. Posteriormente se exponen las características de las técnicas automatizadas empleadas para la segmentación de documentos audiovisuales, incidiendo especialmente en el análisis de vídeos informativos. Para finalizar, se plantean las tendencias que marcan la evolución hacia el análisis semántico de imágenes en movimiento.

2. Segmentación manual de vídeos informativos

El análisis de los vídeos informativos en televisión tiene como objetivo permitir que se recuperen no solo programas completos sino partes de estos que constituyen fragmentos (noticias, secuencias, tomas) potencialmente útiles para una producción o visualización posterior. Tanto la descripción del contenido como la indización tendrán que representar distintos niveles del discurso audiovisual, que constituyen microdiscursos que tienen sentido como partes de una macroestructura jerarquizada y como unidades de significado con valor autónomo. Segmentar un vídeo, determinando cuáles son los microdiscursos que constituyen unidades documentales, es una compleja tarea habitual para los profesionales de la documentación en los medios (López de Quintana, 2007). Antes de generalizarse el acceso “a la carta” el principal objetivo de este análisis era proporcionar imágenes a los periodistas para que los utilizaran en la producción de programas. Actualmente, su función es también facilitar la consulta de los usuarios finales que buscan información puntual sobre un acontecimiento, persona, lugar o noticia concreta.

Aunque los profesionales de las cadenas de televisión cuentan con experiencia suficiente para determinar cuáles son los parámetros para identificar los segmentos de un clip de vídeo que se corresponderían con una unidad de análisis, no se trata de una tarea sencilla ni para la que existan estándares predeterminados. El análisis de las estructuras del discurso audiovisual permite crear patrones que ayudan a identificar unidades significativas independientes. Caldera (2005) señalaba los siguientes tipos de unidades documentales en televisión:

· *Unidad documental de emisión*. Corresponde con las macroestructuras globales de los espacios emitidos (informativo o programa).

· *Unidad conceptual*. Correspondería con las estructuras globales de la noticia, crónica, acontecimiento o reportaje.

· *Unidad documental temática*. Corresponde al contenido de los discursos más amplios (noticia) y los microdiscursos comunicativos que lo conforman (secuencias o conjuntos de secuencias).

Los informativos presentan una estructura bastante regular que estaría formada por a) la cabecera, en el que se desarrollan las principales noticias; b) sumario, con un recorrido breve y rotulado de las noticias a tratar en el informativo; c) noticias, divididas normalmente por bloques temáticos; y d) cierre, en el que se vuelven a destacar las principales noticias. Descendiendo en el nivel de granularidad, las noticias (microdiscursos de un informativo) son estructuras cuya forma podemos sistematizar en el siguiente modelo: a) locución en estudio, con un breve resumen o introducción por parte del presentador sobre la noticia; b) entradilla de periodista, por medio de directo o falso directo, desde el lugar de los acontecimientos; c) desarrollo de la información, a través de recursos visuales y locución; d) incorporación de elementos de postprodución y/o grafismo, para aclarar la información facilitada; y e) cierre, aunque no siempre aparece, es habitual contar nuevamente con el periodista ofreciendo información no en directo sino incluido en la crónica (Caldera, 2005).

Sin embargo, aunque la estructura de los informativos es similar, no es idéntica en las diferentes cadenas de televisión; bien sea porque la empresa busca un estilo propio que la diferencie dentro de las tendencias informativas o por diferencias culturales. La expuesta por Caldera difiere en algunos detalles de la presentada por Kompatsiaris, Merialdo, y Lian, Shiguo (2012) o Bertini, Bimbo y Pala (2001). Zhai, Yilmaz y Shah (2005) quienes ponen de manifiesto que el carácter menos estructurado de los informativos de la CNN respecto a los de la ABC implicaba dificultades para elaborar un patrón que identificara la aparición de un presentador como una transición entre noticias con objeto de facilitar la segmentación automática.

Además de las diferencias derivadas de los estilos de emisión o de los entornos geográficos-culturales, habría que tener en cuenta factores cronológicos: no son iguales los telediarios de los 80 que los de ahora. Estas modas inciden no solo en la estructuración de las emisiones y de los programas, también lo hacen en la identificación de los elementos principales de la información, dejando de lado la descripción de elementos que, en determinado momento, se estiman superfluos.

La praxis profesional también puede ofrecer pautas de segmentación a partir de los contenidos audiovisuales y no solo de las estructuras de los programas. Una breve clasificación de los elementos por los cuáles el documentalista segmenta los vídeos para su análisis como microdiscurso informativo podría identificar los siguientes criterios:

a) Segmentación por acción. Una forma muy habitual es la segmentación de las imágenes por medio de una acción o un suceso. Se recoge las diferentes

secuencias, independientemente de los posibles cambios de planos. Este microdiscurso seleccionado suele ser útil para los recursos (no para totales), aglutinando todas las imágenes respecto al mismo acontecimiento.

b) Segmentación por seguimiento. Denominamos seguimiento cuando la secuencia o conjunto de secuencias siguen a un personaje. Independientemente de la acción o acciones que pudiera ejecutar. El protagonista de la imagen no es tanto la acción como el personaje en sí que las ejecuta. No sólo está pensado para personas físicas, pueden ser seres de la naturaleza.

c) Segmentación onomástica. Muy utilizada por su lógica es la segmentación para unir las imágenes de una persona física o jurídica. Las imágenes que contienen puede ser tanto recursos como totales, aunque por la importancia declarativa del periodismo actual, los totales suelen ubicarse en otra segmentación distinta y, por lo tanto, con metadatos específicos para su acceso.

d) Segmentación declarativa. En este caso la segmentación del vídeo vendrá determinada por la duración y valía de las declaraciones de un personaje. Dentro de una posible segmentación onomástica ésta sería una especificidad derivada del valor de las palabras del personaje. Puede ser útil tanto para material montado como original (bruto o grabación de cámara). Quede claro que no todo lo que diga un personaje es información relevante, por lo que se hace una selección sobre aquello que se dice, destacando lo más relevante, noticiable y/o impactante.

e) Segmentación geográfica. Los recursos derivados de lugares, zonas geográficas, edificaciones, monumentos, etc. son en si mismos información muy útil para los archivos de televisión. De ahí la necesidad de generar segmentación específica para aquella secuencia o conjunto de secuencias que identifiquen un lugar, independientemente de si es un lugar concreto o un accidente geográfico, sin tener en cuenta -aunque se identifique en el análisis descriptivo de las imágenes- el tipo de plano, incidencia angular, etc.

f) Segmentación cronológica. La segmentación realizada debe identificar y concretar la representación temporal de las imágenes. En resumen, el documentalista diferencia entre las imágenes que sean actuales de aquellas que hayan sido extraídas del archivo. Su aspecto audiovisual suele ser distinto en la mayor parte de los casos, incluso contando con diferencias en la coloración (o imágenes incluso en blanco y negro). En caso de que las imágenes de archivo sean recientes y, por lo tanto, pudieran pasar como actuales, el documentalista tendrá en cuenta los tipos de segmentos analizados con anterioridad para discernir si se identifica dicho microdiscurso por separado o, por el contrario, estaría integrado en otros segmentos.

g) Segmentación por postproducción. Este tipo de segmentación destaca aquella información en forma de gráficos, tablas o cualquier otro elemento de postproducción. Esta información no suele ser útil para utilización futura, ya que los gráficos van variando en su estética y, por su puesto, pierden su vigencia y actualidad. Identificarlo por medio de una segmentación específica del vídeo, sirve para eliminarlo en la descripción de otros materiales válidos.

Evidentemente no se trata de una clasificación cerrada y en muchos casos pueden solaparse algunos grupos. Sin embargo, la elaboración de patrones que justifiquen la segmentación de vídeos es, como veremos, una de las tendencias de investigación de la segmentación automática de vídeo, si esta se quiere hacer por contenidos y no solo por elementos formales.

3. Segmentación de vídeo automática.

De forma genérica la segmentación automática consiste en la división o extracción de partes de un vídeo digital mediante un algoritmo o programa informático (San Andrés; Chávez, 2013). Al igual que hemos visto en el apartado anterior, un vídeo puede ser analizado/segmentado a diferentes niveles de granularidad que representan diferentes niveles del discurso audiovisual.

El nivel más elemental es el de la imagen/fotograma, generalmente utilizado para extraer características visuales como color, textura o forma. El siguiente nivel está representado por las tomas[5], según Hu, et al. (2011) una toma es el conjunto consecutivo de imágenes capturado por una cámara entre las operaciones de inicio y fin de la grabación. Smeaton (2006) indica que puede incluir movimientos de cámara como zoom, panorámica, travelling, etc., así como movimientos de objetos. Sin embargo, este no es un nivel relevante para representar partes pertinentes de un programa porque normalmente dura solo pocos segundos y tiene poco contenido semántico. Se necesitan técnicas de "alto nivel" para agrupar las tomas en segmentos más significativos: las secuencias. Una secuencia se compone generalmente de un grupo de tomas relacionadas con la misma materia, desarrollo de un evento o tema. La ambigüedad que implica la interpretación de esta definición se extiende a las denominaciones que se han utilizado para designarla: escenas, párrafos de vídeo, unidades de historia, capítulos, etc. (Abduraman, Berriani y Merialdo, 2012)

Detección de tomas: métodos basados en características visuales

Las técnicas automáticas de detección de tomas son las más estudiadas porque permiten trabajar esencialmente con características visuales. Si trasladamos la definición de tomas a términos técnicos, estas deberían considerarse como un grupo de imágenes (frames) que comparten características visuales (color, textura, movimiento). Normalmente, la dirección de la cámara y el punto de vista definen una toma: cuando una cámara reproduce la misma escena desde diferentes ángulos, o diferentes partes de la misma escena desde distintos ángulos, estaremos ante diferentes tomas. Como las tomas se caracterizan por la coherencia de las características visuales de "bajo nivel", segmentar los vídeos en tomas resulta una

[5] Traducimos de esta manera el término inglés shot, por considerar que es el término español más acorde con la definición. En otras obras también ha sido traducido como plano temporal, etc.

tarea relativamente fácil (Ogawa, et al., 2008). Por otra parte, se considera que las tomas son las unidades fundamentales para organizar el contenido de las secuencias de vídeo y la base (primitives) para la anotación semántica y las tareas de recuperación de más nivel (Hu, et al., 2011).

En los vídeos editados, las tomas están separadas por transiciones que pueden ser abruptas (cortes) o graduales (separación mediante algún efecto de edición: disolver, barrido, desvanecerse, etc.). En este caso, la identificación de los límites de una toma consiste en reconocer las transiciones, abruptas o graduales, entre tomas adyacentes. Por su simplicidad, fue el primero que se utilizó, ya que el corte del vídeo venía determinado por un corte en el montaje. Este método, que fue muy utilizado en los departamentos de las cadenas de televisión, resulta bastante fácil de implementar computacionalmente. No obstante, esta fórmula presenta problemas importantes ya que los cambios de toma graduales, los más utilizados en televisión, plantea graves problemas. Los cambios de planos abruptos no suelen ser utilizados por su falta de estética, por lo que suelen utilizarse fundidos más o menos rápidos, cortinillas, transiciones entre secuencias que en muchos casos son identificados por la máquina como un gran número de tomas distintas, segmentando un vídeo (en el caso de fundido) en más de siete veces para menos de medio segundo (Angulo, 1999).

Resultan más difíciles de implementar los métodos para la detección de los límites de las tomas que solo utilizan las características visuales de las imágenes para calcular su similitud, agrupando aquellas que comparten atributos similares (video mining). En este caso, los límites de las tomas se sitúan entre fotogramas que son disimilares.

Entre las características para la detección de los límites de una toma incluyen:

Segmentación por color. Esta fórmula de identificación de elementos tiene muchas ventajas derivadas de que el análisis viene determinado por colores o gamas de colores semejantes tal como se reflejan en los histogramas o bloques de color. La máquina aprecia que el objeto (independientemente de su naturaleza) es el mismo si los colores siguen constantes (San Andrés, Franco; Chávez, 2013). Los histogramas de color son robustos para pequeños movimientos de cámara, pero no permiten diferenciar las tomas dentro de una misma escena y son sensibles a movimientos de cámara grandes.

Existen dos fórmulas para realizar esta segmentación: apagando colores o encendiendo colores. Esto implica que el cambio podrá realizarse tanto por la falta de colores existentes en los anteriores frames o por la adición de otros. Al ser la gama de colores muy amplia y entendiendo el análisis desde un punto de vista de economía de proceso (velocidad de procesamiento y consumo de rendimiento) solo se trabaja con los colores rojo, verde y azul. El algoritmo utilizado mapea la imagen en celdas, y siempre que detecta un color primario en una imagen, analizará los posibles cambios que pudiera haber en dicha celda. Cuando los cambios son significativos realizará la segmentación de vídeo.

Segmentación por bordes (formas). Se basa en la detección de los bordes por medio de la detección de valores de intensidad de las zonas. Algunos algoritmos suavizan la imagen en primer lugar, reduciendo “ruido” audiovisual (detalles y

texturas). Posteriormente, la imagen se transforma a escala de grises, para que los colores no puedan confundir y queden claros los bordes. En las imágenes se identifican los píxeles en los que se produce la máxima variación, eliminando de la ecuación aquellos que varían de forma poco significativa (eliminación de falsos bordes). Las características de los bordes son menos variables a los cambios de iluminación y al movimiento que los histogramas.

*Segmentación por textura***.** Es el método por el cual se distinguen las diferentes parcelas de la imagen por su textura. Por textura nos referimos a la aspereza, iluminación, rugosidad, espacios libres, sombras, etc. En definitiva, la segmentación por textura analiza los valores detectados por el análisis de los brillos y la escala de grises. Al igual que la segmentación por bordes, la segmentación por textura requiere pasar la imagen a escala de grises para aplicar el filtro de aplicación. En cada píxel se analiza la intensidad de la luz, al igual que se analiza cada píxel en relación con los de su entorno. De esta manera se marca la textura y la segmentación de regiones en la imagen. Cuando estas texturas y brillos cambian de forma significativa se lleva a cabo la segmentación de vídeo.

Segmentación por modelado de fondo. Esta técnica utiliza elementos, algoritmos y parámetros de algunos de los anteriores modelos (Herrero, 2009). Este modelo consiste en algoritmos que mantienen un modelo matemático de apariencia del fondo, para así poder detectar los objetos en movimiento que están en primer plano y aquellos que están en el fondo. Resulta un método complejo debido a que requiere el reconocimiento objetos en movimiento (vectores de movimiento), lo que es posible por: a) el análisis de cambios de iluminación que nunca debieran interpretarse como objetos móviles; b) análisis de sombras y reflejos; c) la actualización constante del fondo de la escena; d) eliminación de "ruidos" audiovisuales del fondo captados en el momento de la grabación de las imágenes; e) detección de fondos en movimiento constante (movimiento del agua, copas de los árboles, etc.; y f) detección del camuflaje, que consiste en diferenciar los elementos del primer plano respecto del fondo, aunque tenga el mismo color y/o textura.

Medidas de similitud

El paso siguiente para identificar los límites de las tomas consiste en medir la similitud entre fotogramas utilizando las características extraídas. Algunas de estas medidas incluyen la similitud del coseno, la distancia euclídes, la intersección de histogramas o la similitud chi-cuadrado. Pueden aplicarse entre fotogramas consecutivos (medidas inter-pares) o entre un conjunto de fotogramas (vidriera). Los métodos de detección de los límites pueden clasificarse en:

- basados en un umbral que detectan los límites comparando las similitudes entre pares de fotogramas con un umbral determinado. El umbral puede ser global, adaptable (flexible) o una combinación de ambos. Los algoritmos globales usan el mismo umbral, que generalmente se determina de forma empírica teniendo en cuenta la información de todo el vídeo. Los algoritmos flexibles computan el umbral localmente dentro de una ventana móvil.

- basados en aprendizaje estadístico, que consideran la identificación de límites como una tarea de clasificación, en la que se clasifican los fotogramas como cambio de toma/no cambio de toma dependiendo de las características que contienen. Pueden ser supervisados o no supervisados.

A su vez, las diferentes opciones de detección de límites pueden clasificarse dentro del ámbito de la compresión o fuera de este ámbito. Para evitar el consumo de tiempo del video no comprimido pueden emplearse las características empleadas en el dominio comprimido. Sin embargo, la aproximación de vídeo comprimido es muy dependiente de los estándares de compresión y es menos precisa que la del vídeo sin comprimir.

Una correcta segmentación de vídeo ayudará a una generación de keyframes identificativos del contenido del documento de calidad y fiables, lo cual genera la oportunidad de no tener que acceder al documento completo, sino analizar por medio de estas imágenes fijas la validez del material (Vilches, 2001). La detección de estas imágenes significativas constituye un gran valor para la rapidez de acceso y como resumen de contenido. Este sistema de identificación de key-frames extrae las imágenes por medio de similitud entre ellas, por lo que la calidad vendrá determinada por una correcta segmentación de la información audiovisual.

Identificación de secuencias (escenas): la segmentación semántica de vídeo.

Los primeros trabajos sobre segmentación se centraron en la utilización de la información visual. Sin embargo, los trabajos posteriores reconocieron la importancia del mensaje de audio en la segmentación de vídeo centrándose en la detección de secuencias (Ogawa et al.). La segmentación de escenas también se conoce como "story unit segmentation". En general, una escena es un grupo de tomas contiguas que son coherentes con determinada materia o tema. Las escenas tienen un nivel semántico más alto que las tomas y se identifican o segmentan agrupando tomas sucesivas con un contenido similar dentro de una unidad semánticamente significativa. El objetivo es identificar estructuras que permitan establecer patrones, aplicando técnicas de vídeo "structure mining" (Vijayakumar, Nedunchezhian, 2012).

La agrupación de tomas de vídeo en escenas depende de juicios subjetivos sobre su correlación semántica. Estrictamente hablando, tal agrupación requiere la comprensión del contenido semántico del vídeo. Sin embargo, uniendo características de audio y vídeo, a veces es posible reconocer escenas que están relacionadas por su localización o por el tipo de evento, sin que esto implique un análisis de contenidos semánticos.

Estas técnicas pueden dividirse en dos grandes categorías: métodos genéricos y métodos específicos. Los métodos específicos explotan el conocimiento previo en un domino para construir un modelo estructurado del vídeo analizado. Solo pueden aplicarse a tipos de programas muy específicos como noticias, deportes, series, publicidad, etc. Los métodos genéricos buscan una aproximación universal para estructurar vídeos, independientemente de su tipo y basada solo en las características de su contenido. Los autores que proponen métodos específicos

consideran que es muy difícil, si no imposible, conseguir una solución universal para analizar los vídeos a nivel semántico (alto nivel). Las características de nivel bajo generalmente utilizadas para indizar contenidos de vídeo no son suficientes para proporcionar información semánticamente significativa, por lo que es necesario utilizar información contextual (audio, texto o conocimiento del dominio).

De acuerdo con la representación de tomas, los métodos de segmentación de escenas pueden clasificarse en tres categorías (Hu et al., 2011):

Basadas en Keyframes: se extraen las características de los keyframes que representan cada toma y se agrupan en una escena aquellas que presentan características similares. La limitación de este método radica en que los keyframes pueden no representar con eficacia los contenidos dinámicos de las tomas.

Basadas en fondos: Este método segmenta las escenas asumiendo que las tomas que pertenecen a la misma escena frecuentemente tienen fondos similares. Esta es, a su vez, su mayor limitación ya que a veces los fondos de las tomas que pertenecen a la misma escena tienen fondos diferentes.

Basadas en la integración de información audio y visual: este método selecciona una escena cuando coinciden los límites de una toma y un cambio en la banda de audio. Su limitación es que es difícil determinar la relación entre segmentos de audio y tomas.

De acuerdo con los métodos de procesamiento, las tendencias actuales en segmentación de escenas pueden dividirse en cuatro categorías (Hu, et al., 2011):

a) Aglomerativas, en las que gradualmente se agrupan tomas similares hasta formar una escena (botton-up style).

b) Divisivas, estilo top-down, se parte de todo el clip de vídeo y se van identificando escenas.

c) Modelos estadísticos, se construyen modelos estadísticos de las tomas para segmentar las escenas.

d) Basadas en la clasificación de los límites de las tomas, se utilizan las características de los límites de las tomas y se utilizan para clasificar los límites de las tomas en límites de escenas y no límites de escenas.

El punto en común de estos cuatro métodos es que se utilizan las similitudes entre tomas para combinar tomas similares en escenas. Esto es simple e intuitivo. Sin embargo en estas aproximaciones las tomas generalmente se representan por un conjunto de keyframes, que frecuentemente no representan los contenidos dinámicos de las tomas. En consecuencia, dos tomas se consideran similares si sus keyframes están en el mismo entorno más que si son similarmente visuales. La que tiene en cuenta los límites de las tomas se beneficia de la información local sobre esos límites, lo que permite utilizar algoritmos poco complejos.

Los textos que pueden ayudar en el análisis automático del vídeo pueden obtenerse gracias a las técnicas de reconocimiento automático de voz, de los textos de los subtítulos de determinados programas o por la identificación de texto en las imágenes gracias a técnicas de OCR. “Físchlár-News video library system” es un buen ejemplo en el que la búsqueda y la navegación se construyen a partir del

diálogo hablado (Smeaton, 2006).

El término características semánticas en el contexto de la recuperación de vídeo significa la identificación de características de nivel medio o alto que expresan contenido semántico. Algunos ejemplos pueden incluir interior/exterior, vegetación natural, personas, personajes conocidos, carreteras, aviones. Estas son características mucho más difíciles de detectar automáticamente que las características de bajo nivel, como colores y texturas, que pueden extraerse sin dificultad.

Un elemento crítico en la detección de estas características es tener acceso a un amplio corpus de vídeo con "anotaciones" manuales de alta calidad que permita entrenar un sistema informático para realizar una clasificación basada en criterios semánticos. En este caso el trabajo previo debe realizarse manualmente por lo que puede ser factible para un número limitado de conceptos, pero no lo es cuando se trata de los miles de conceptos que se manejan en ámbitos generalistas como son los informativos.

Sin embargo, la mayoría de las aproximaciones explotan las características de dominios específicos como películas, deportes e informativos. La segmentación de emisiones de noticias en escenas fue una de las tareas más importantes de evaluación en TRECVid 2003 y 2004 (Smeaton, 2006). La descripción de la tarea define escena como "segmento de un informativo centrado de forma coherente en una noticia (tema, evento) y que contiene al menos dos clausulas declarativas" (Misra, et al., 2010). Para hacerlo se han estudiado métodos que utilizan modelos basados en las reglas de producción para crear modelos a priori y técnicas que añaden texto, audio y vídeo a las características visuales.

En lo referente a la segmentación de textos, existen métodos que analizan la repetición de palabras claves para identificar el límite de cada historia. Para ello es necesario la transcripción de la información, asociando términos como "buenos días", "hasta mañana", "devolvemos la conexión", etc.

La combinación de las diferentes realidades (textual –incluidos subtítulos y rotulación-, sonora y visual) aportan según Hsu, et al. (2004) los mejores resultados.

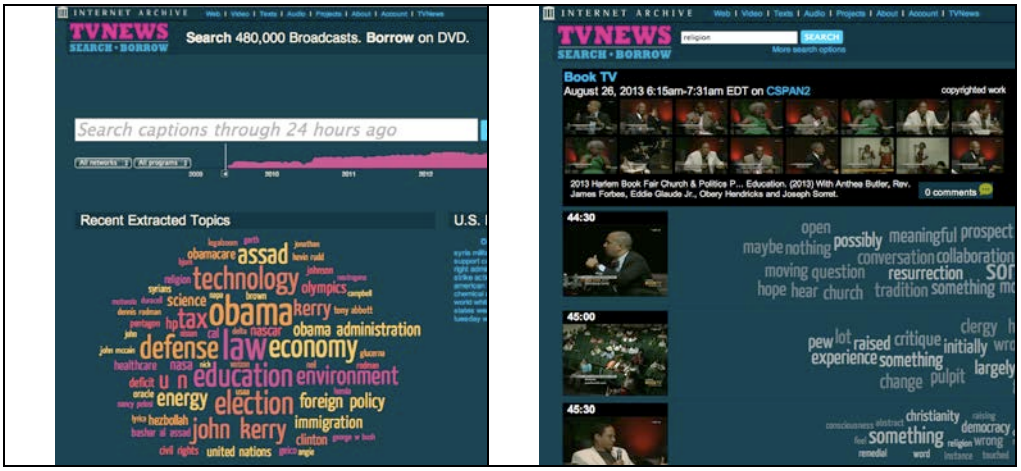

Figura 2. Internet Archive. Video. News. 1 Búsqueda titulares de actualidad 2. Sumario magacine informativo y palabras clave de los shots. Fuente: http://archive.org/details/tv

El trabajo relacionado con el análisis de la estructura de programas de noticias es mucho más extenso y anterior que el de programas de deportes. Esto puede deberse parcialmente a la regularidad de su estructura, lo que hace el análisis mucho más sencillo. Coincidiendo con la estructura básica planteada por Caldera (2005),

Adburaman, Berrani y Merialdo (2012) señalan que la mayoría de los informativos tienen una estructura regular y bien definida que generalmente sigue este esquema: vista general del plató con el presentador anunciando el sumario, y el resto del programa organizado como una sucesión de noticias e imágenes del presentador anunciando el siguiente tema. Cada noticia empieza con imágenes del presentador que proporciona una descripción general del evento y continúa con un reportaje más detallado y, a veces, entrevistas. Al final de una historia un segmento con el presentador puede hacer un sumario o conclusión. La mayoría de las emisiones acaban con reportajes sobre noticias, tiempo y finanzas. Durante la emisión también pueden aparecer anuncios publicitarios.

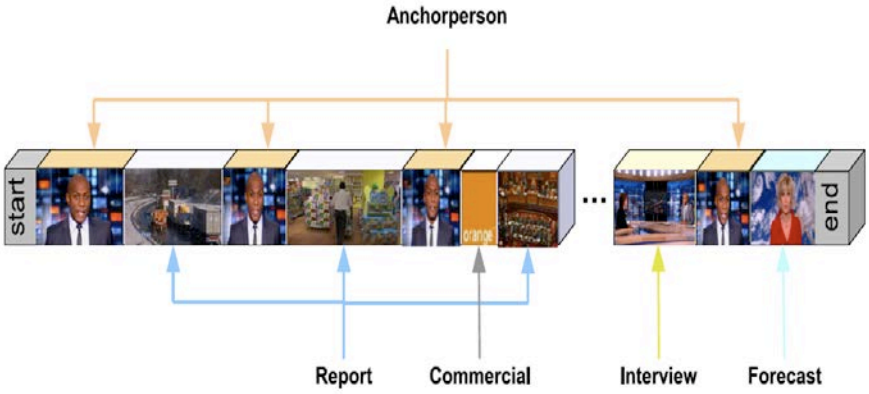

Figure 1.2: The structure of a TV news program.

Figura 3. Estructura de programas informativo de televisión
Fuente.
https://www.researchgate.net/publication/276928907_TV_Program_Structuring_Techniques_A_Review

Aplicando el método basado en patrones, el primer paso para reconstruir la estructura de un programa de noticias es clasificar las tomas en diferentes grupos (presentador, noticias, predicción del tiempo, entrevistas, etc.) para pasar después a hacer la segmentación del vídeo en unidades-historia (noticias/escenas). La detección de las tomas con presentadores juega un papel muy importante en estos métodos porque son fáciles de reconocer mediante técnicas de equiparación basadas en medidas de similitud e imágenes modelo, utilizando técnicas de reconocimiento facial (Correira, et al, 2004). También se aprovechan características de producción como el hecho de que el fondo apenas cambie durante la toma, que se filmen habitualmente con cámaras estáticas que siempre están situadas en el mismo lugar respecto a la imagen, que haya menos movimiento que en las tomas de noticias o la presencia de objetos representativos a nivel semántico como pueden ser los logos o leyendas con el nombre del presentador, que se visualizan mientras los presentadores hablan (Bertini et al., ; Caldera y Zapico, 2009; Gunsel, et al., 1998; Zhang et al., 1995).

4. Conclusiones

Llevar a cabo una segmentación de vídeo previa a la labor humana ayuda en gran manera al trabajo posterior. De todas maneras, se podrá retocar y volver a unir clips de vídeo y dividir aquellos que se considere oportuno, pero existe una clasificación previa que pudiera servir de base para la confección de la labor documental.

Vilches (2001) trata el concepto de segmentación semántica, de manera sucinta, para identificar la necesidad de mejorar la segmentación de vídeo para adaptarla a los requerimientos reales, aportando mejoras por medio del reconocimiento de rostros. En la actualidad no puede entenderse la segmentación y su análisis desde un punto de vista semántico, ya que lo que se realiza son aproximaciones al contenido de la información mediante el análisis de los objetos visualizados, por medio de la parametrización de elementos como el color, la textura, el movimiento, etc.

Nos referimos a segmentación semántica al hecho de fragmentar los vídeos atendiendo a pautas y patrones de requerimientos humanos, los cuales hemos descrito con anterioridad. Es decir, unir los algoritmos matemáticos de la inteligencia artificial a las posibilidades de la biométrica, para así encaminarnos hacia la segmentación por elementos coherentes para los documentalistas y los usuarios (periodistas para el caso de la televisión). Junto a parámetros de segmentación de vídeo automático señalados, habría que utilizar técnicas de reconocimiento facial, de tal manera que además de segmentar podría identificar a personajes, lo cual sería especialmente útil para las televisiones (Caldera; Zapico, 2009). Igualmente podría utilizar otros parámetros para la identificación de objetos por medio de la biometría, lo que supone generar grandes bases de datos con modelos para que la máquina pueda comparar. Dichos modelos pueden ser de, por ejemplo, seres vivos así como de edificios. Esto nos ayuda no sólo a realizar la segmentación por muchos de los parámetros señalados por los documentalistas, sino que además ayudará a la automatización del proceso de descripción de imágenes.

Existen otras herramientas que contribuyen a la mejora de la segmentación, como pudiera ser el reconocimiento de voz. Los personajes que aparecen en televisión son un número muy limitado, y los que hablan son aún menos. Por lo tanto, los reconocedores de voz identificarán los personajes que hablan en las imágenes, identificando al personaje por un lado y ayudando a determinar el principio y fin de las imágenes atendiendo a lo escuchado. Podríamos plantearnos que aquellas personas anónimas no serán identificadas por esta herramienta. Esto no es problema, ya que estos personajes no serán realmente válidos para su reutilización futura. Además, podría realizarse lectura por medio de OCR para identificar los textos que aparecen en los rótulos de postproducción por el cual se identifican a los personajes implicados en la información.

Estas son líneas computacionales de trabajo para segmentar correctamente las imágenes, para que dicha segmentación sea realmente útil para los documentalistas. Debe adaptarse la tecnología existente para acercar la segmentación manual del documentalista a la segmentación automática, lo cual potenciará una futura automatización del proceso de indización y descripción de imagen, y la extracción de keyframes representativos del contenido con mucha mayor calidad.

5. Reconocimiento

Este trabajo ha sido financiado por el Gobierno de Extremadura (Consejería de Educación, Ciencia y Tecnología) y el Fondo Social Europeo dentro del plan de apoyo a las actuaciones de los Grupos de Investigación inscritos en el catálogo de la Junta de Extremadura. GR10019.

6. Referencias

Abduraman, Alina Elma; Berrani, Sid Ahmed; Merialdo, Bernard (2012). TV program stucturing tecniques. En: *TV Content Analysis: Techniques and Applications*. Kompatsiaris, Yiannis ; Merialdo, Bernard ; Lian, Shiguo (eds.). Boca Raton, FL : CRC Press. ISBN 978-1-4398-5560-7

Angulo López, Jesús (1999). *Segmentación temporal de secuencias de vídeo*. Valencia: Universidad Politécnica. Proyecto fin de carrera.

Bertini, M.; Del Bimbo, A.; Pala, P. (2001). Content-based indexing and retrieval of TV news. *Patter Recognition Letters*, 22 : 503-516.

Bilge Gunsel, Ahmet Mufit Ferman, A. Murat Tekalp, "Temporal video segmentation using unsupervised clustering and semantic object tracking," Journal of Electronic Imaging 7 (3) : 1-32.

Caldera-Serrano, J (2005). Unidades semánticas discursivas en la información televisiva. *Ciencias de la Información*, vol. 36, n. 3, pp. 39-48

Caldera-Serrano, J (2006). Terminological control of “anonymous groups” for catalogues of audiovisual television documents. *Journal of Librarianship and Information Science*, 38 (3), 187-195.

Caldera-Serrano, J; Zapico-Alonso, F (2009). Reconocimiento facial biométrico para identificación onomástica en archivos de televisión. *El Profesional de la Información*, vol. 18, n. 4, pp. 427-431.

Correia, P.L.; Pereira, F.M. (2004) Classification of Video Segmentation Application Scenarios. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, n. 5, pp. 735-741

Herrero Martín, Sonsoles (2009). *Análisis comparativo de técnicas de segmentación de secuencias de vídeo basadas en el modelado del fondo*. Madrid: Universidad Autónoma. Trabajo fin de carrera.

Gunsel, B.; Ferman, A.M.; Tekalp, A.M. (1998). Temporal video segmentation using unsupervised clustering and semantic object tracking. Journal of Electronic Imaging, 73: 1-32

Hsu, W.H., Kennedy, L.S., Chang, S.-F., Franz, M., Smith, J.R.(2004). *Columbia-IBM. News Video Story Segmentation in TRECVID 2004*. In: TREC.

Hu, W.; Xie, N.; Li, L.; Zeng, X.; Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. *IEEE Trans Syst Man Cybern C: Appl Rev* , pp. 1–23.

Kompatsiaris, Y.; Merialdo, B.; Lian, S. (2012). TV content analysis: techniques and applications. EE.UU.: CRC Press.

López de Quintana, Eugenio (2007). Transición y tendencias de la documentación en televisión: digitalización y nuevo mercado audiovisual. *El Profesional de la Información*, vol. 16, n. 5, pp. 397-408.

Misra, H.; Hopfgartner, F.; Goyal, A. et al. (2010). Tv news story segmentation based on semantic coherence and content similarity. *Proceedings of the 16th international conference on Advances in Multimedia Modeling*, pp. 347–357.

Ogawa, A.; Takahashi, T.; Ide, I.; Murase, H. (2008). Cross-lingual retrieval of identical news events by near-duplicate video segment detection. En: *Advances in Multimedia Modeling: 14th International Multimedia Modeling Conference*, MMM 2008, Kyoto, Japan, January 9-11, 2008.kioto: Springer, pp.287-296. ISBN: 978-3-540-77407-5

San Andrés Lascano, Xavier Ernesto; Franco Pombo, Vicente Julian; Chávez Burbano, Patricia (2011). *Implementación de Interfaz Gráfica para Comparación Visual de Métodos de Segmentación y Procesamiento de Vídeo, Usando Matlab*.

http://www.dspace.espol.edu.ec/handle/123456789/19032

Smeaton, Alan F. (2007). Techniques Used and Open Challenges to the Analysis, Indexing and Retrieval of Digital Video. Information Systems, 32 (4): 545-559.

Vilches, Lorenzo (2001). Tecnologías digitales al servicio de los archivos de imágenes. *Anàlisi*, vol. 27, pp. 133-150.

Vijayakumar, V.; Nedunchezhian, R. (2012). A study on video data mining. *Int J Multimed Info Retr*, 1, pp.153–172

Zhai, Yun ; Yilmaz, Alper ; Shah, Mubarak (2005). Story segmentation in news videos using visual and text cues. En: *CIVR'05: Proceedings of the 4th international conference on Image and Video Retrieval*: Singapore, July 20-22, pp. 92-102 (También disponible en http://dpl.ceegs.ohio-state.edu/papers/yilmaz_civr_2005.pdf)

Zhang, H.; Furth, B.; Smoliar, S.W. (1995). *Video and imagen processing in multimedia systems*. Kluwer, Dordrecht.