DETECCIÓN DE CAMBIOS DE TOMA CON INFORMACIÓN DE CONTENIDO VISUAL Y AUDITIVO

Alejandro Abejon e Ismael Mateos[1]

ATVS (Biometric Recognition Group), C/ Francisco Tomas y Valiente 11,
Universidad Autonoma de Madrid, E28049 Madrid, Spain
{alejandro.abejon, ismael.mateos}@uam.es

RESUMEN

En este artículo se aborda la detección de cambios de toma en contenidos audiovisuales desde dos perspectivas complementarias, la información existente en la parte visual y la información de audio. Adicionalmente se proponen varios métodos de combinación de estas informaciones para obtener un sistema final robusto. La extracción de características de la parte visual se realiza a través de los descriptores *GoF/GoP* y *Scalable Color* de MPEG-7. Para la extracción de la información de la parte de audio se emplea BIC *(Bayesian Information Criterion)*. En la sección experimental se demuestra como la técnica de audio presenta una alta precisión en la detección, mientras que la técnica visual muestra un nivel de *recall* elevado. La combinación de ambas técnicas mejora sensiblemente el comportamiento final de la detección.

Palabras claves: detección de cambios de toma, MPEG-7, *GoF/GoP*, *Scalable Color,* BIC.

1. INTRODUCCIÓN

Durante los últimos años se ha experimentado un gran aumento en la cantidad de contenidos audiovisuales. Este aumento ha venido propiciado principalmente por dos causas: en primer lugar la constante aparición de nuevos dispositivos cada vez más baratos y con mejores prestaciones, en segundo lugar el desarrollo de estándares (compresión, codificación, etc.) que hacen cada vez más sencilla y eficiente la distribución de contenidos.

La anotación automática es un proceso importante a la hora de tratar con grandes bases de datos de contenidos audiovisuales (clasificación, almacenamiento, distribución, etc.). Este conjunto de técnicas, también conocido como servicios de valor añadido [1, 2], enriquecen la información multimedia de cara a procesos de búsqueda, indexación y auto-resúmenes.

Todos estos sistemas tiene en común la necesidad de aplicar una segmentación temporal de los contenidos para poder realizar una anotación automática de los mismos. Existen dos conceptos clave en este proceso: por un lado la segmentación debe ser lo mejor posible, por otro lado la velocidad debe ser elevada, ya que es necesario que la segmentación sea más rápida que tiempo real.

Una de las formas más comunes de segmentar contenido audiovisual consiste en la detección de cambios de toma. Esta tarea ha sido enfocada tradicionalmente desde el punto de vista de la señal de video [3]. Aunque hay algunos autores como Shu-Ching Chen en 2002 y Yingying Zhu y Dongru Zhou en el año 2003 [4, 5], que emplearon una combinación de la información obtenida del audio y la imagen de forma que el sistema contara con una mayor robustez. El problema principal de estas técnicas es la combinación de la información de audio y video [6]. Se han propuesto varias soluciones para dicha combinación de información [7, 8] sin llegar a obtener un resultado plenamente satisfactorio.

La estructura que se va a seguir en este artículo es la siguiente. En primer lugar, en la sección 2 se realiza un recorrido por el estado del arte en técnicas de detección de cambios de toma con información visual, información de audio y posibles combinaciones de ambas. Posteriormente, en la sección 3 se detalla el sistema implementado, la técnica empleada a nivel visual, a nivel de audio y las combinaciones propuestas. Los resultados obtenidos se muestran en la sección 4. Por último, la sección 5 contiene las conclusiones extraídas de la realización de este trabajo.

2. ESTADO DEL ARTE

2.1. Detección a nivel visual

La detección de cambios de toma basada en información visual es la aproximación más utilizada. Podemos distinguir dos tipos de cambios de toma: los cambios abruptos en los que el cambio se da de un frame al siguiente; por otro lado están los cambios graduales cuya duración es de varios frames. Los cambios abruptos son los más fáciles de detectar, en los últimos años se han propuesto varios algoritmos que gozan de una gran precisión [9, 10]. Por el contrario, la complejidad en la detección de cambios graduales es mayor. En la literatura podemos distinguir dos vertientes: la que afronta cierto tipo de cambios específicos (ej. disoluciones) con resultados aceptables [11], por otro lado, podemos encontrar técnicas que tratan de construir modelos generales que engloben cualquier tipo de cambio gradual posible [12, 13]. Esta última aproximación presenta una fiabilidad menor que la anterior.

El histograma de color es una de las variables más comúnmente empleadas en sistemas visuales de detección de cambios de toma. Este hecho es debido a que un cambio de toma suele llevar consigo un cambio en las distribuciones de color de la imagen [14]. Otro factor que se observa cuando hay un cambio de toma es el incremento del número y módulo de los vectores de movimiento [15]. Codificaciones como MPEG-2 o H.264 se basan en la similitud de imágenes sucesivas para codificar bloques en función de bloques de imágenes adyacentes. Un cambio de toma llevará consigo un

[1] Este trabajo ha sido apoyado por el Ministerio de Ciencia y Educación, TEC2006-13170-C02-01. Los autores desean mostrar su agradecimiento al grupo GTI de la Escuela Politécnica Superior (UAM) por su apoyo.

incremento en el número de vectores de movimiento, debido al menor parecido entre imágenes.

En el proceso de la detección del cambio de toma podemos distinguir dos fases: medida de disparidad entre parejas de frames, que pueden estar consecutivas o distanciadas y toma de decisión de si esa disparidad entre frames es o no un cambio de toma.

2.2. Detección a nivel de audio

La tarea de detección de cambios en la señal de voz está englobada dentro de lo que se conoce como *anotación de un fichero de audio*. El objetivo de estas técnicas es enriquecer cualquier contenido de audio, dotándole de una cierta información acerca de locutores, música, silencios, ruido, etc. que en el fichero aparezcan.

Existen distintos niveles a la hora de anotar un fichero de audio, desde el más sencillo que consistiría simplemente en indicar los tramos del fichero de voz y no voz (dentro de la categoría de no voz se incluye música, silencio, ruido, etc.), hasta los más elaborados en los que se indican los tramos donde está presente el mismo locutor, tramos con música, silencio, etc.

Una hipótesis de partida podría ser que cambios de toma en la señal visual llevan asociados cambios en la señal de audio. Por tanto el objetivo será detectar variaciones en la señal sonora: como por ejemplo el paso de voz a música. Otra alternativa sería tener en cuenta los cambios de audio como puntos clave a la hora de segmentar, sin necesidad de que lleven asociados cambios de toma.

Para llevar a cabo la detección de cambios de toma existen principalmente dos fases: la primera de ellas consistente en la detección de voz y silencio, la segunda está relacionada con la detección de tramos con música, ruido, diferentes locutores, etc. dentro de los segmentos de voz. Para la primera de las etapas se suelen emplear los cruces por cero de la señal aunque la aproximación más habitual está basada en el cálculo de la energía. Para llevar a cabo la segunda etapa existen dos técnicas principalmente, ambas hacen uso de ventanas que se van desplazando a lo largo del fichero de audio.

El criterio de información bayesiano *(BIC)* [16] es una de las dos aproximaciones para la detección de distintos tramos dentro de segmentos de audio. Esta técnica busca puntos de cambio dentro de una ventana, para ello se comprueba si los datos de la ventana se modelan mejor con una única distribución (no hay punto de cambio) o con dos distribuciones (punto de cambio). Si se encuentra un punto de cambio se resetea la ventana y se reinicia la búsqueda a partir de ese punto. Si por el contrario no se encuentra punto de cambio se incrementa la ventana y se vuelve a realizar la búsqueda. La búsqueda completa a lo largo del fichero es muy costosa computacinalmente, del orden de N^2 siendo N el número de muestras del fichero, por tanto la mayor parte de los sistemas emplean versiones simplificadas de este algoritmo.

La segunda técnica empleada se propuso en 1997 [17], consiste en usar ventanas de una longitud fija (normalmente entre 2 y 5 segundos) que serán representadas por una gausiana. Posteriormente se calculan las distancias entre ventanas. Viendo si estas distancias superan o no un umbral seremos capaces de encontrar los puntos de cambio. La longitud de las ventanas limita la detección de cambios de corta duración.

En los últimos tiempos han aparecido otras técnicas de mayor complejidad aunque su comportamiento es comparable al de las técnicas anteriores. Dichas técnicas utilizan modelos de mezclas de gausianas *(GMM o Gaussian Mixture Models)* o modelos ocultos de Markov *(HMM o Hidden Markov Models)* para modelar música, locutores y ruido ambiental. El gran inconveniente de estas técnicas es que son supervisadas y para su entrenamiento se necesita una gran cantidad de datos etiquetados que modelen la generalidad de aquello que se quiere detectar.

2.3. Combinación de información de audio y video

La combinación de la información procedente del audio y del vídeo es una de las labores más complejas en un sistema de detección de cambios de toma. Aunque la información de audio va ligada a la información visual no tienen porque coincidir con los cambios de toma reales. Puede darse el caso de que se produzca un cambio en el audio, como por ejemplo la intervención de un nuevo locutor, que no lleve consigo un cambio de toma y viceversa.

Son pocos los trabajos encontrados en la literatura que traten este problema. En [4] se hace una primera división basada en la señal de audio que divide la señal en silencio, voz y música; dentro de la parte de voz se hace distinción entre locutores. Por otro lado se realiza la segmentación basada en el contenido visual, para posteriormente combinar ambas informaciones dando un mayor peso a la señal de audio.

Existen otras posibilidades a la hora de realizar la fusión entre las informaciones de audio y vídeo. Desde los trabajos en los que no se realiza ninguna fusión [8], pasando por trabajos en los que se realiza una segmentación a nivel de audio y después se usa la información visual para comprobar si es correcto [18], hasta variantes de la aproximación anterior donde se confirman los cambios extraídos del vídeo con el audio [5].

3. SISTEMA IMPLEMENTADO

3.1. Segmentación a nivel visual

La técnica implementada realiza una detección de tomas mediante un modelo general, es decir, lo que buscamos es que dicho modelo detecte tanto los cambios de toma abruptos como los graduales. Para ello nos basaremos en dos de los descriptores definidos por MPEG-7: *GoF/GoP* y *Scalable color* (más información en: ISO/IEC 15938-3 6.5, ISO/IEC 15938-3 6.8, ISO/IEC TR 15938-8). El objetivo original de estos descriptores es la búsqueda de similitudes entre vídeos y/o imágenes.

Este sistema es equivalente a realizar una detección de cambio de toma por histograma de color, ya que los coeficientes sobre los que se aplican las medidas de dispersión miden la distribución de color en cada frame analizado.

Las medidas de dispersión evaluarán el grado de similitud entre frames, dichas medidas se aplican sobre los coeficientes resultantes del descriptor. En lugar de calcular los coeficientes para todos las frames se realiza para una de cada 25 frames, es decir un frame por segundo. Con este procedimiento conseguimos dos objetivos, por una lado reducir la carga computacional y por otro la detección de cambios graduales. Como medidas de dispersión se emplearon 3 de las que mejor comportamiento presentaban en [14].

$$\delta_1 = \sum\sum \left| c_a(x,y) - c_b(x,y) \right|^2 \quad (1)$$

$$\delta_2 = 1 - \min\left(LR, \frac{1}{LR} \right) \quad (2)$$

$$LR = \frac{\left[\frac{S_a^2 + S_b^2}{2} + \left(\frac{c_a - c_b}{2} \right)^2 \right]^2}{S_a^2 + S_b^2} \quad (3)$$

donde *s* es la dispersión y *c* es la media de cada trama.

La tercera de las medidas de dispersión es una combinación de las dos anteriores, propuesta también en [14]. Consiste en aplicar un filtro de mediana de longitud tres, a las dos medidas de dispersión anteriores y restarlas a las originales. Esta técnica proporciona una mayor robustez frente a flashes y cambios de iluminación. Una vez tenemos calculados el vector diferencia realizamos el calculo del modulo de dicho vector.

$$\lambda = \left\| (\delta_1, \delta_2) - mediana(\delta_1, \delta_2) \right\| \quad (4)$$

Nuestra propuesta consiste en establecer el umbral de forma dinámica calculando el 20% del margen dinámico de cada una de las tres medidas indicadas. Siempre que las tres componentes superen este umbral se decidirá que existe un cambio de toma. Esta forma de establecer el umbral ofrece una mayor robustez, pero tiene el inconveniente de tener que recorrer el fichero antes de la fase de toma de decisión.

3.2. Segmentación a nivel de audio

El proceso seguido para obtener la información de audio consta de dos pasos básicamente: *i)* una vez separado el contenido de audio del vídeo se realiza una parametrización MFCC *(Mel Frequency Cepstral Coefficients)* de 13 coeficientes; *ii)* a continuación se aplica el algoritmo de detección de cambios, BIC (véase sección 2.2).

3.3. Combinación de la información visual y de audio

Antes de llevar a cabo la combinación de la información obtenida del audio y del vídeo se realiza un filtrado de los resultados para subsanar en lo posible errores en la detección. Este filtrado consiste en eliminar cambios de toma sucesivos que tengan entre sí menos de un segundo de duración. Se presupone que un cambio de toma por muy rápido que sea debe durar más de un segundo, por lo que los cambios de este tipo que resulten de nuestros algoritmos serán eliminados.

Una vez realizado este filtrado estamos en disposición de combinar la información de los dos métodos para obtener un resultado global más robusto. Como se observa en el estado del arte, esta combinación no resulta sencilla. En nuestro caso se propone dos tipos de combinaciones distintas.

En primer lugar, la más restrictiva de ellas (combinación &), consiste en indicar sólo aquellos cambios de toma en los que la información obtenida del audio y del vídeo coincida. Con el fin de dotar de mayor flexibilidad al sistema introduciremos un cierto margen de error programable que para las pruebas que mostraremos en la sección 4 será de 3 segundos. El sistema dictaminará que hay un cambio de toma siempre que vídeo y audio coincidan con un margen de error de más menos 3 segundos.

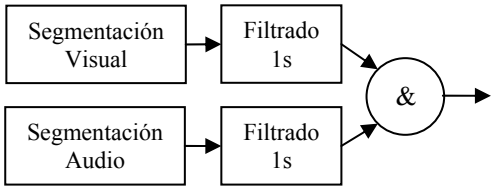

Figura 1. *Esquema combinación restrictiva segmentaciones, comb. &*

En segundo lugar emplearemos como método de combinación algo bastante sencillo (combinación +), la agrupación de los cambios indicados por cada uno de los métodos por separado. Una vez hecho esto se realizará el proceso de filtrado mencionado anteriormente, pero con menos de tres segundos de duración.

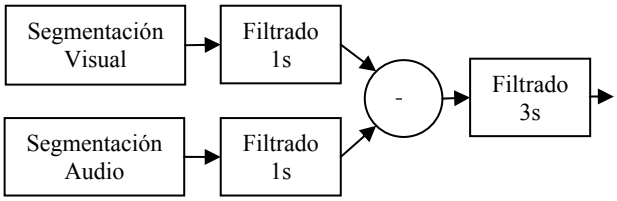

Figura 2. *Esquema combinación concatenada segmentaciones, comb. +*

4. RESULTADOS

El vídeo de ejemplo seleccionado cosiste en 5 minutos (7500 frames), extraídos del programa *Informe Semanal*. Este tipo de vídeo se puede considerar dentro de la categoría de informativos, una de las más estudiadas a lo largo de los últimos años. El video a nivel visual contiene 7 cambios graduales y 22 cambios abruptos. La señal de audio consta de 17 cambios, no todos ellos coincidentes con el vídeo y por tanto no detectables por técnicas basadas en audio.

La Tabla 1 muestra los resultados obtenidos con la detección visual por un lado y la detección a través de audio por otro. Además se muestran estos resultados tras el filtrado y las combinaciones de ambas técnicas propuestas.

	OK	NO	FP	R (%)	P (%)
Visual	10	19	19	34	34
Visual filtrado	9	20	7	31	56
Audio	7	22	1	24	88
Audio filtrado	7	22	1	24	88
Comb. &	3	27	1	11	75
Comb. +	11	18	10	38	52

Tabla 1. *Resultados de detección de cambios de tomas; OK cambios correctos, NO cambios no detectados, FP falsos positivos, R recall, P precesión*

A la vista de los resultados vemos como el filtrado aumenta la precisión del sistema, sobre todo en la parte visual. En la parte de audio no se observa cambio alguno ya que no se detectaban cambios tan seguidos. Vale la pena destacar que cuando la detección es solo con audio, la precisión que se obtiene es de las más altas, lo que nos indica que los cambios detectados con el audio tienen una alta probabilidad de ser correctos. Además todas medidas están calculadas sobre el número total de cambios a nivel visual, por tanto la técnica a nivel de audio está en clara desventaja.

También se observa que la primera combinación adoptada (comb. &), aquella que implica que el cambio se dé tanto en la parte visual como en la de audio es demasiado restrictiva, lo cual provoca que a penas se detecten cambios de toma. Por el contrario cuando simplemente unimos los cambios detectados en el audio y en la imagen (comb, +) se observa que la cantidad de cambios de toma correctos es elevada. Las Figuras 3 y 4 muestran los cambios de toma detectados con las dos combinaciones propuestas.

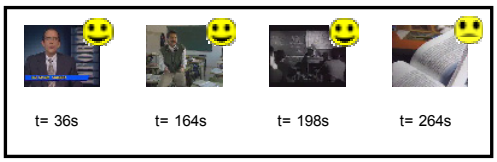

Figura 3. *Resultados combinación de segmentaciones &*

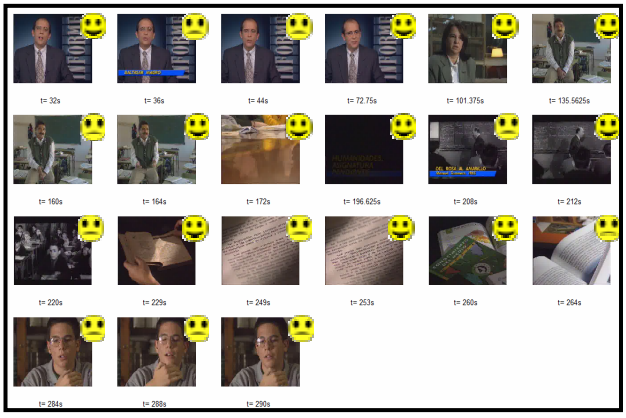

Figura 4. *Resultados combinación de segmentaciones +*

5. CONCLUSIONES

En este artículo se ha abordado uno de los problemas más habituales en la anotación de contenidos audiovisuales: la detección de cambios de toma. Para llevar a cabo esta labor se emplean dos tipos de informaciones: *i)* la información de audio y *ii)* la información de vídeo. Ambas informaciones son combinadas con el objetivo de obtener un sistema global más robusto. La información de cambios de audio se extrae mediante BIC. La información del contenido visual se obtiene siguiendo las indicaciones propuestas por MPEG-7 en dos de sus descriptores: *GoF/GoP* y *Scalable Color*.

La sección de resultados muestra como la información de audio es importante y ayuda a la mejora del rendimiento del sistema global. La precisión obtenida a través del contenido de audio es de las más elevadas. Este hecho arroja resultados esperanzadores y anima a la búsqueda de métodos de combinación más eficientes. Por otro lado, una alternativa a la detección de cambios de toma de cara a la realización de auto-resúmenes sería la segmentación del contenido basada únicamente en la información del audio. La señal de audio contiene información suficiente como para poder llevar a cabo esta tarea sin apoyarse en la información visual.

Dos aspectos importantes de la implementación son en primer lugar la velocidad del análisis, siendo para el audio muy eficiente. En segundo lugar, la parte visual está basada en descriptores escalables, este hecho hace que tanto la velocidad en la extracción de coeficientes como la precisión sean configurables y adaptables a las necesidades de cada sistema.

6. BIBLIOGRAFÍA

[1] MERL Video Summarization for PVRs (2008) http://www.merl.com/projects/VideoSummarization [Online].

[2] Lire (Lucene Image REtrieval) (Junio 2008) http://www.semanticmetadata.net [Online].

[3] Koprinska and S. Carrato, "Temporal video segmentation, a survey," *Signal Processing. Image Commun*., vol. 16, no. 5, pp. 450-477, Jan. 2001.

[4] S. Chen, et al., "Scene Change Detection by Audio and Viedo Clues", *IEEE Transactions on Speech and Audio Processing*, Vol. II, pp. 365- 368, 2002.

[5] Y. Zhu and D. Zhou, "Scene Change Detection Based on Audio and Video Content Analysis", *ICCIMA 2003.*

[6] A. Yoshitaka, and M. Miyake, "Scene Detection by Audio-Visual Features," *IEEE International Conference on Mulatimedia and Expo (ICME01)*, pp. 49-52, 2001.

[7] H. Sundaram and S.-F. Chang, "Video Scene Segmentation Using Video and Audio Features," *IEEE International Conference on Multimedia and Expo (ICME00)*, pp. 1145-1148, 2000.

[8] T. Muramoto and M. Sugiyama, "Visual and Audio Segmentation for video streams," *IEEE International Conference on Mulatimedia and Expo (ICME00),* pp 1547-1550, 2000.

[9] L. Qianlei, et al., "Twi-difference Algorithm and Pixel-matching Twi-difference Algorithm for Video Abrupt Shot Change Detection",*Journal of Image and Graphics* A, Vol. 2, No. 2, pp. 161-168, 2003.

[10] S. Gong and Y. Fan, "Video abrupt shot change detection based on relation of the partial interframe differences," *Machine Learning and Cybernetics*,. Volume 9, Page(s):5255-5260 Vol. 9, Aug. 2005.

[11] C. Su et al., "A motion-tolerant dissolve detection algorithm", *IEEE Transactions Multimedia,* Volume 7, Issue 6, Page(s):1106 – 1113 Dec. 2005.

[12] W. Xiong and J. C. M. Lee, "Efficient scene change detection and camera motion annotation for video classification," *Comput. Vis. Image Understanding*, vol. 71, no. 2, pp. 166–181, Aug. 1998.

[13] P. Bouthemy, et al., "A unified approach to shot change detection and camera motion characterization," *IEEE Trans. Circuits Syst. Video Technol*., vol. 9, pp. 1030–1044, 1999.

[14] J. Bescós, "Real-Time Shot Change Detection Over Online MPEG-2 Vedeo", *IEEE Trans. Circuits Syst. Video Technol* Vol 14 No 4 pp 475-484 April 2004.

[15] M. Zhi and A. Cai, "Shot change detection with adaptive thresholds",*VLSI Design and Video Technology*, Page(s):147 – 149 May 2005.

[16] S. S. Chen and P. S. Gopalakrishnam, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, 1998, pp. 127–132.

[17] M. A. Siegler, et al.,"Automatic segmentation, classification and clustering of broadcast news," in *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, Feb. 1997, pp. 97–99.

[18] H. Jiang, et. al., "Video Segmentation with the assistance of audio content analysis," *IEEE International Conference on Mulatimedia and Expo*, pp. 1507-1510, 2000.