A Fast and Effective Method for Static Video Summarization on Compressed Domain

A. C. Hernandez, M. C. Hernandez, F. G. Ugalde, M. N. Miyatake and H. P. Meana

[1]***Abstract*— Current advances in information technologies have led to the creation of huge video databases. Efficient and effective video summarization methods are needed to quickly browse and locate required video content. In this paper, we present a fast and effective video summarization method that is implemented in the compressed domain. Our four-step proposed method relies on a simple yet powerful descriptor and a scene-detection method, to detect gradual and abrupt transitions with great precision. A saliency-based refinement strategy is used to avoid redundancy and represent video content with as few key-frames as possible. Several experiments were done to assess the proposed method's performance, concluding that it is superior to current solutions.**

***Keywords*— Video Summarization, Key-Frame Extraction, Compressed domain**

I. Introducción

La producción y el consumo de videos en formato digital se han incrementado de manera considerable durante la última década. Esto ha significado la generación de enormes bases de datos de video que precisan de mecanismos eficientes de gestión para recuperar contenido de forma rápida y eficaz [1], [2]. Las técnicas de resumen de video permiten eliminar la redundancia existente en una secuencia de video para generar una representación abreviada de su contenido. De esta forma, se agiliza la navegación en las grandes bases de datos ya que se tiene una comprensión inmediata de los videos, en sentido semántico [3]-[6]. De acuerdo a los elementos audiovisuales que proveen al usuario, las técnicas de resumen de video se pueden clasificar en técnicas de *cuadros clave*, que presentan imágenes representativas del video en orden temporal, y técnicas de *segmentos de video*, las cuales crean fragmentos breves de video para detallar su contenido [7]. Las técnicas de cuadros claves tienen ventajas significativas, como lo son un bajo costo computacional, una mayor rapidez para precisar el contenido de un video y la posibilidad de ser un insumo para aplicaciones avanzadas de análisis de video [8], [9]. Además, las técnicas de cuadros clave son aplicadas bajo dos enfoques. Las técnicas de *representación global*, conciben el video como un conjunto de imágenes sin orden específico que se agrupan de acuerdo a su similitud [10]-[12]. Algunas de las técnicas de agrupamiento empleadas en la literatura científica actual, son el método de K-medias [10], la triangulación de Delaunay [11] y el agrupamiento basado en la densidad de aplicaciones con ruido (DBSCAN) [12]. Estos métodos requieren de un elevado costo computacional debido a que: a) necesitan decodificar la secuencia de video por completo, y b) llevan a cabo procesos adicionales para conocer con anticipación el número total de agrupaciones [10], reducir los datos a procesar, o reordenar los cuadros de video para generar resúmenes más coherentes [11]. Por otro lado, el enfoque de *representación local* propone una división del video en escenas a partir del análisis de cuadros sucesivos, para después elegir un cuadro de cada una de ellas. Estos métodos pueden generar resúmenes de forma progresiva debido a su bajo costo computacional [13]. No obstante, tienen problemas para detectar cambios de escena cuando estos son graduales y no abruptos, y generan duplicidad en el resumen final cuando una escena se repite a lo largo del video.

En este trabajo de investigación se presenta un método de resumen de video simple, ágil, y efectivo, el cual se basa en la extracción de cuadros clave con un enfoque de representación local. Conociendo los inconvenientes asociados a este tipo de métodos, se proponen procedimientos de aplicación práctica para dar solución a cada uno de ellos. Nuestro método tiene por objetivo eliminar la redundancia temporal para elegir el menor número de cuadros posible que representen mejor el contenido del video. Además, el método tiene muy bajo costo computacional debido a que utiliza datos obtenidos a partir de la trama codificada de video. A continuación, se enlistan las contribuciones científicas más relevantes de este trabajo:

- La precisión del método para clasificar escenas se basa en un descriptor de color simple pero eficaz que es una propuesta innovadora de los autores.
- El método propuesto realiza un procesamiento único y secuencial de los cuadros de video, lo cual permite que el usuario obtenga un resumen progresivo sin esperar a que el video se procese por completo.
- De esta forma, se consigue un modelo eficiente que no precisa de tareas de refinación adicionales al final del proceso de resumen, como sucede en otros métodos.
- Adicionalmente, se realiza una aportación para mejorar el mecanismo de evaluación de resúmenes de video.

Los resultados obtenidos muestran que el método propuesto genera resúmenes de video con mayor calidad en comparación a cuatro destacados métodos de la literatura científica actual.

II. Método propuesto

A. Segmentación temporal

La primera etapa del método de resumen propuesto consiste en una segmentación temporal de la secuencia de video. Esto implica extraer sólo la información más importante de la trama de video para establecer las unidades básicas que darán

A. C. Hernandez, Universidad Nacional Autonoma de Mexico, Ciudad de Mexico, Mexico, antoniochz@hotmail.com
M. C. Hernandez, Instituto Politecnico Nacional, Ciudad de Mexico, Mexico, mcedillohdz@hotmail.com
F. G. Ugalde, Universidad Nacional Autonoma de Mexico, Ciudad de Mexico, Mexico, fgarciau@unam.mx
M. N. Miyatake, Instituto Politecnico Nacional, Ciudad de Mexico, Mexico, mnakano@ipn.mx
H. P. Meana, Instituto Politecnico Nacional, Ciudad de Mexico, Mexico, hmperezm@ipn.mx
Corresponding author: Antonio Cedillo Hernandez

inicio al proceso. En el contexto de los estándares MPEG-1/2/4 los cuadros inter-codificados (I) son la opción más adecuada para realizar la segmentación inicial, debido a que: a) representan el contenido visual de todos los cuadros que forman un grupo de imágenes (GOP), b) cada GOP se reduce al análisis de un cuadro I, lo cual genera una reducción de 15 veces de los datos a procesar, y c) decodificar sólo la información de los cuadros I permite ahorrar hasta un 80% del costo computacional total requerido para decodificar la secuencia de video [14]-[15].

Además, para ahorrar costo computacional adicional, cada cuadro I se procesa para crear una versión reducida del mismo. Un cuadro I se divide en bloques de 8×8 bits transformados a patrones de frecuencia por medio de la transformada discreta de coseno (DCT). Cada bloque DCT tiene un componente DC, cuyo valor denota la luminancia promedio y 63 componentes AC de media y alta frecuencia. En el método propuesto, cada cuadro I se representa por su versión reducida, conocida como cuadro-DC, que resulta al excluir los componentes AC de cada bloque DCT [16]. Esta estrategia aporta algunas ventajas: a) se procesa sólo 1/64 de la información original, b) el cuadro-DC contiene la información espacial de la imagen, y c) a pesar de tener baja resolución, el cuadro-DC sólo reduce el 7% de la capacidad de un humano para distinguir detalles visuales [17].

B. Extracción de características

Posteriormente, cada cuadro-DC se procesa para calcular su descriptor de color utilizando los datos de crominancia sin la necesidad de cambiar de espacio de color, lo cual es común en algunos métodos. El procedimiento para calcular el descriptor propuesto es el siguiente (Fig. 1): 1) Aislar las componentes de crominancia (Cr, Cb) de cada cuadro-DC. 2) Generar un histograma bidimensional (H_c), que muestre la ocurrencia de valores de crominancia en un rango no signado (0-255), donde el eje x de H_c denota los valores de la componente Cr, y el eje y los de la componente Cb. 3) Crear una máscara binaria (B) a partir de H_c mediante una operación binaria, definida como:

$$B(x,y)=\begin{cases}1, & H_c(x,y)>0\\ 0, & \text{otro caso}\end{cases} \quad \text{donde } x,y=0,\ldots,255 \tag{1}$$

4) Basados en un concepto de firma de imagen, y con la meta de asociar una firma única a cada cuadro de video, se diseña una plantilla de mezcla de texturas, denotada como T, con las mismas dimensiones que H_c (256×256 pixeles). Inicialmente, la plantilla T se divide en cuatro cuadrantes T_i de dimensiones n×n, donde i=1,2,3,4 y n=128. A cada cuadrante T_i se le asigna diferente contenido visual a partir de imágenes de referencia en escala de grises que cumplen las siguientes condiciones: a) Dada una imagen de referencia en escala de grises, se obtiene su densidad óptica promedio (AOD) [18], mediante (2):

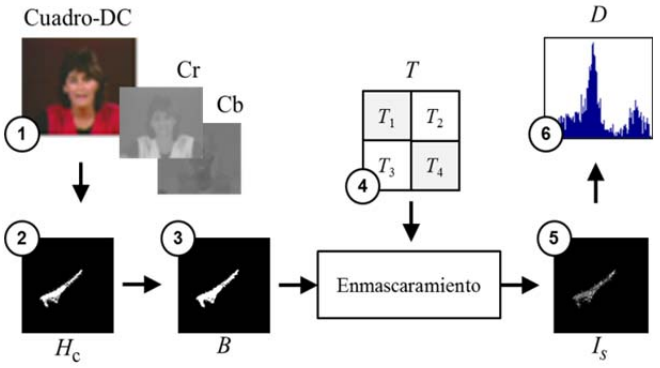

Figura 1. Proceso para calcular el descriptor de color propuesto.

TABLA I
CONDICIONES AOD Y CONTENIDO VISUAL POR CUADRANTE

T_i	AOD	Contenido visual
1	≈0.56	Textura débil y contenido plano dominante
2	≈0.51	Texturas fuertes dominantes
3	≈0.44	Balance entre texturas fuertes y débiles
4	≈0.53	Textura débil y contenido plano dominante

$$AOD(f)=\frac{1}{N_1N_2}\sum_{k=0}^{K-1}k\cdot H_f(k) \tag{2}$$

donde f es la imagen a procesar de dimensiones N_1×N_2, H_f es el histograma de f, K es la saturación máxima en la escala de gris con un valor igual a 255, y k = 0,..,K-1. b) Se define un criterio de contenido visual que está determinado por la cantidad de textura que contienen las imágenes de referencia, utilizando el algoritmo de clasificación reportado en [19]. Así, se elige una imagen de referencia para cada cuadrante T_i de acuerdo a los valores de AOD y el contenido visual, mostrados en la Tabla I. El método puede operar con diversas imágenes de referencia, siempre que cumplan con los requisitos establecidos. 5) Se realiza una operación de enmascaramiento con los datos de B y T para calcular una firma denotada como I_s (3):

$$I_s(x,y)=\begin{cases}T(x,y), & B(x,y)=1\\ 0, & \text{otro caso}\end{cases} \quad \text{donde } x,y=0,\ldots255 \tag{3}$$

6) Por último, se calcula el histograma unidimensional a partir de la firma I_s, para calcular el descriptor de color basado en la firma de imagen, el cual se representa como D.

C. Detección de escenas

La siguiente etapa del método propuesto consiste en dividir el video en escenas. Para ello, se comparan los descriptores de cuadros consecutivos a través de la norma L_2 (4) [20]:

$$d(D_t,D_{t-1})=\left(\sum_{r=1}^{N}\left|\left(D_t^r-D_{t-1}^r\right)^2\right|\right)^{1/2} \tag{4}$$

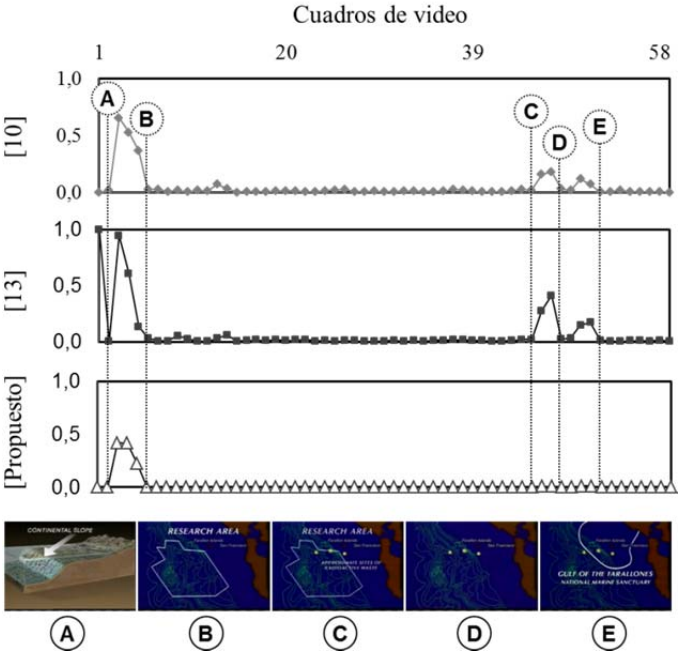

Figura 2. Las fluctuaciones visuales entre los primeros 60 cuadros de video de la secuencia *Ocean Floor Legacy*, *segment 4*.

donde D^r es el *r*-ésimo elemento del descriptor D, D_t y $D_{t\text{-}1}$ son los descriptores en los tiempos t y t-1 respectivamente y N es la longitud de D. La Fig. 2 muestra el resultado obtenido al realizar la comparación de los primeros sesenta cuadros de la secuencia de video *Ocean Floor Legacy, segment 4*, utilizando el método propuesto y los métodos presentados en [10] y [13]. La Fig. 2 muestra la precisión de cada método para detectar cambios abruptos (A-B) y cambios moderados de contenido (C-D-E). A partir de la Fig. 2, podemos determinar que los tres métodos detectan correctamente los cambios abruptos, sin embargo sólo el método propuesto obtiene valores cercanos a cero cuando los cambios no son significativos, lo cual genera mayor precisión al momento de detectar cambios de escena.

Como se mencionó antes, las técnicas de cuadros clave con un enfoque de representación local, presentan problemas para detectar cambios graduales de escena, debido al bajo contraste existente entre cuadros sucesivos [21], [22]. Para resolver este inconveniente, se usa el método de *comparación doble* [23], el cual se basa en definir dos umbrales para detectar escenas, uno para cambios abruptos (T_h) y otro para cambios graduales (T_l). Cuando se manifiesta un cambio gradual (T_l), las diferencias sucesivas se acumulan en espera de que eventualmente se supere el valor de T_h. Cuando esto sucede, el cuadro inicial es marcado como el final de la escena anterior (E_F), y el cuadro final de la transición se marca como el inicio de la siguiente escena (E_I). Cuando un cambio abrupto es detectado (T_h), los cuadros anterior y posterior se marcan como E_F y E_I. La Fig. 3 ilustra gráficamente el método de *comparación doble*.

D. Extracción de cuadros clave

En la última etapa del método propuesto se elige el cuadro ubicado al centro de cada escena (E_I - E_F), como cuadro clave para formar parte del resumen final. Este cuadro suele resumir los elementos visuales de una escena de forma adecuada.

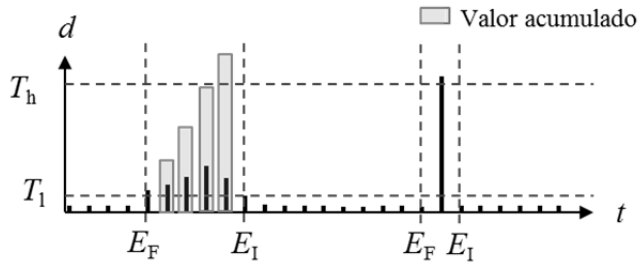

Figura 3. Método de *comparación doble* [23].

Sin embargo, cuando se elige un cuadro clave, no se tienen elementos para determinar si su contenido es similar al de otro cuadro que ya es parte del resumen. Este tipo de redundancia se genera cuando un video contiene escenas repetidas que se presentan alternadamente a lo largo del tiempo, por ejemplo, el video de una entrevista. Para evitar este efecto, cada vez que se elige un cuadro clave, éste es procesado para determinar las regiones de saliencia dentro de su contenido. Para esto, se usa el algoritmo propuesto en [24], el cual se define a como sigue:

$$\overline{x} = \mathrm{IDCT}\left[\mathrm{sign}\left(\mathrm{DCT}\left(x\right)\right)\right], \tag{5}$$

$$m = \sigma * \left(\overline{x} \circ \overline{x}\right), \tag{6}$$

donde x es el componente de luminancia del cuadro-DC, m es el mapa de saliencia obtenido a partir del filtro Gaussiano con desviación estándar σ, y el símbolo ○ denota el producto de Hadamard. A partir del mapa de saliencia m, se obtiene su representación binaria (S) con la aplicación de un umbral T_S:

$$S(x,y) = \begin{cases} 1, & m(x,y) \geq T_S \\ 0, & m(x,y) < T_S \end{cases}, \tag{7}$$

Esta representación binaria se utiliza para comparar el cuadro clave actual con aquellos obtenidos previamente, a través de la tasa de error binaria (BER). Con esta medición se determina la tasa de bits idénticos entre dos imágenes binarias. Un cuadro clave se descartará, sólo si el BER obtenido entre su descriptor binario y el de los cuadros clave previos, es mayor a 0.9.

III. Resultados

A. Mecanismo de evaluación

Para evaluar la calidad de los resúmenes de video generados se utiliza una versión modificada del método de *comparación de resúmenes de usuarios* (*CUS*) [10]. Éste método se basa en dos métricas denominadas tasa de precisión (CUS_A) y tasa de error (CUS_E), que miden la distancia entre un resumen creado por un usuario (S), de otro generado de forma automática (A). Si concebimos a un resumen de video como un conjunto de cuadros clave, la comparación de los conjuntos S y A tiene tres posibles resultados, los cuales se representan como r_1, r_2 y r_3 en la Fig. 4. La tasa de precisión se define como la cantidad de cuadros coincidentes entre S y A (r_1), entre el número total de cuadros del conjunto S (n).

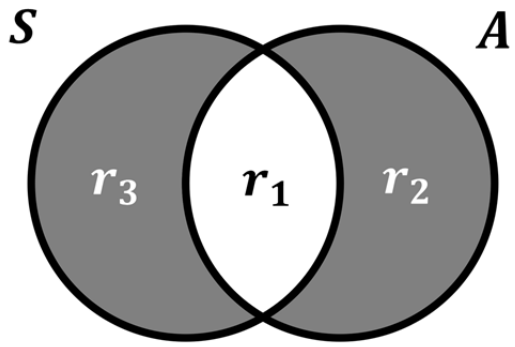

Figura 4. Subconjuntos r_1, r_2 y r_3 resultantes de la comparación de un resumen generado por un usuario (S) y otro calculado automáticamente (A).

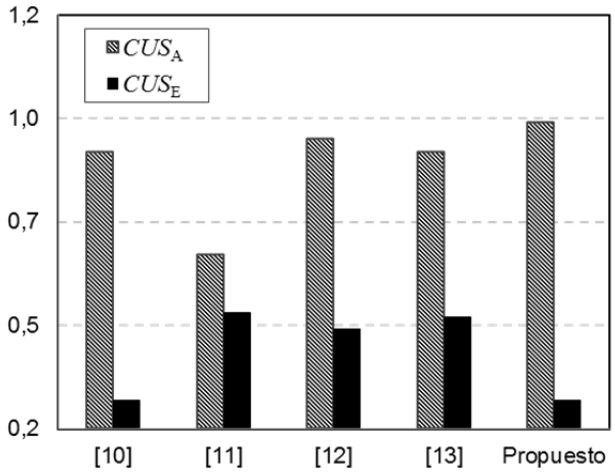

Figura 5. Resultados promedio obtenidos para CUS_A y CUS_E por cada uno de los métodos evaluados.

Por otra parte, la tasa de error se define como la cantidad de cuadros que se incluyen en A, pero que el usuario no consideró en su resumen S (r_2), divididos entre n. Se puede afirmar que el subconjunto r_2 contiene los cuadros *sobrantes*, ya que son cuadros que el usuario no deseaba dentro de su resumen. Con este enfoque, la métrica estaría ignorando los cuadros que el usuario consideró en su resumen S y que no se incorporaron automáticamente en A (r_3), es decir, los cuadros *faltantes*. Para obtener una medición más precisa, se propone redefinir la tasa de error de forma que se consideren los subconjuntos r_2 y r_3. De esta forma, ambas métricas son definidas de la siguiente forma:

$$CUS_A = \frac{r_1}{n} = \frac{|S \cap A|}{|S|} \tag{8}$$

$$CUS_E = \frac{r_2 + r_3}{n} = \frac{|S \triangle A|}{|S|} \tag{9}$$

B. Desempeño del método propuesto

Para evaluar el desempeño del método propuesto, se generó una base de datos con 50 videos descargados del sitio web de *Open Video Project*, los cuales están codificados en formato MPEG-1, con una resolución espacial de 352×240 pixeles, 30 FPS y una duración promedio de 2 minutos.

Los experimentos reportados se obtuvieron a partir de un prototipo desarrollado bajo la plataforma MATLAB ®. Los umbrales T_h y T_l fueron calculados de forma experimental con valores de 0.45 y 0.11, respectivamente. El umbral T_S se obtiene mediante el método global de Otsu [25].

Los resultados obtenidos se comparan con cuatro métodos de la literatura científica actual [10]-[13], los cuales tienen un desempeño sobresaliente. Afortunadamente, se puede realizar una comparación justa debido a que los métodos en [10]-[12] usan la misma base de datos de prueba que se utiliza en este trabajo. Además, los resultados reportados para [10]-[12] se calculan a partir de los resúmenes de video publicados por los autores a través de los sitios web de cada proyecto. En el caso de [13], se realizó una simulación computacional a partir del método publicado. Los resúmenes de video generados por los usuarios se obtuvieron a través del sitio web de la propuesta en [10], los cuales se encuentran disponibles en línea [26].

La Fig. 5 muestra los resultados promedio obtenidos por cada uno de los métodos evaluados. El mejor resultado para la precisión (CUS_A) es obtenido por el método propuesto, lo que indica que en la mayoría de los casos, el conjunto de cuadros clave obtenidos por el método propuesto corresponden con los cuadros elegidos por los usuarios. En valor más óptimo para la tasa de error (CUS_A) es obtenido por el método propuesto y el método en [10], sin embargo la precisión de éste método es significativamente menor. Para conocer con mayor detalle los resultados obtenidos por este trabajo, se ha publicado un sitio web con los resultados de rendimiento obtenidos y el conjunto de cuadros clave generado para cada secuencia de video, así como la base de datos de prueba [26].

La Fig. 6 muestra el resumen de video generado por el usuario, los métodos [10]-[13] y el método propuesto, a partir de la secuencia *Drift Ice as a Geologic Agent, segment 8*. En la Fig. 6 podemos observar que el más bajo desempeño es obtenido por el método [11]. En general y debido a su diseño, este método genera resúmenes muy cortos que no contienen en su mayoría, todos los cuadros de video elegidos por el usuario. Los métodos en [10] y [12] tienen mejores resultados que [11] debido a que sus resúmenes son más grandes y parecidos a los aportados por los usuarios. El problema con estos dos métodos es que son realizados con un enfoque de representación global lo que indica que su costo computacional es mayor. Por último, el método en [13] tiene un enfoque de representación local con la posibilidad de crear resúmenes progresivamente al igual que el método propuesto. A pesar de que su tasa de precisión es alta, este método suele elevar su tasa de error ya que presenta el inconveniente clásico de no poder detectar cambios de escena graduales con mucha precisión.

C. Costo computacional

En esta sección se analiza el costo computacional requerido por el método propuesto y los métodos en [10]-[13]. La Tabla II muestra el resultado de éste análisis, utilizando la notación asintótica. En la Tabla II, n denota la cantidad de cuadros necesarios para ejecutar cada uno de los métodos, y d el tamaño del vector que describe las características de cada cuadro de video.

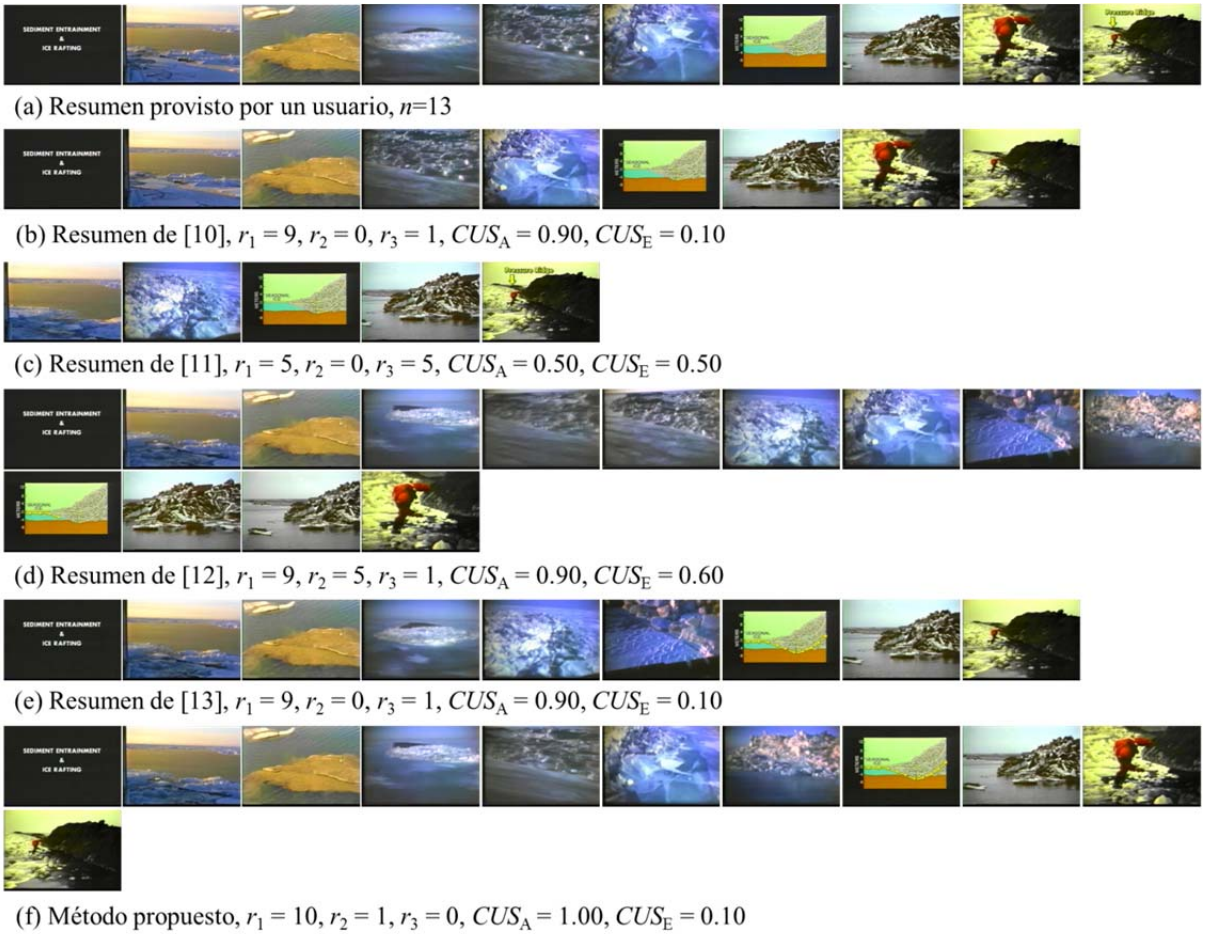

Figura 6. Resultados promedio obtenidos para CUS_A y CUS_E por cada uno de los métodos evaluados.

TABLA II
ANÁLISIS DE COSTO COMPUTACIONAL

Método	Almacenamiento	Costo computacional
[10]	$O(n \cdot d)$	$O(k \cdot n \cdot 2^k + n)$
[11]	$O(n \cdot d)$	$O(n \log n)$
[12]	$O(n \cdot d)$	$O(n \log n)$
[13]	$O(d)$	$O(n)$
Propuesto	$O(d)$	$O(n)$

Podemos observar que las técnicas de representación global [10]-[12] requieren almacenar todos los descriptores de la secuencia de video para llevar a cabo el proceso. En cambio, los métodos [13] y propuesto, solo precisan de almacenar el descriptor actual para funcionar adecuadamente. En el caso del costo computacional, el método en [10] requiere de un costo exponencial que depende del número de agrupamientos k [28], el método en [11] requiere de un costo logarítmico debido al método de agrupamiento DBSCAN. El costo del método [12], es analizado en [29], concluyendo que requiere de un costo logarítmico. Al final, queda en evidencia que los métodos de representación local ([13] y propuesto) tienen un costo mucho menor que los de representación global. Sin embargo, a pesar de que el método propuesto tiene un costo computacional que es similar al presentado en [13], éste presenta notables mejoras de desempeño, lo cual lo hace un método más eficiente.

IV. CONCLUSIONES

Las grandes bases de datos requieren métodos automáticos que permitan administrar su contenido para realizar consultas rápidas y eficientes. Dado que los videos son almacenados de forma codificada, resulta impráctico decodificar cada trama de video para llevar a cabo su análisis. El método de resumen de video que se presenta en este trabajo es una solución altamente aplicable en escenarios prácticos ya que trabaja directamente con videos codificados. Diversas técnicas son aplicadas para eliminar la redundancia espacial y temporal, y de esta forma generar resúmenes de video con el menor número de cuadros clave posible. Los autores proponen un descriptor de color que en conjunto con una estrategia de detección de escenas, ha demostrado su eficacia para detectar transiciones graduales y abruptas. Además, una técnica basada en un mapa de saliencia es empleada para mejorar la calidad de los resúmenes creados. Los resultados experimentales, basados en una medición que es una versión mejorada del método *CUS*, demuestran que el método propuesto puede generar resúmenes de calidad con un costo computacional muy bajo.

AGRADECIMIENTOS

Este trabajo fue realizado gracias a los apoyos recibidos por parte del programa de becas posdoctorales de la Universidad Nacional Autónoma de México (UNAM), el Consejo Nacional de Ciencia y Tecnología y el Instituto Politécnico Nacional.

REFERENCIAS

[1] R. Martin et al., "neXtream: A multi-device, social approach to video content consumption," in *Proc. IEEE CCNC*, Jan. 2010, pp. 1–5.

[2] F. Pereira. "Video Compression: An Evolving Technology for Better User Experiences" *II National Conference on Telecommunications*, Keynote Lecture 5. May 2011.

[3] A. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *J. Vis. Commun. Image Represent*, vol. 19, no. 2, pp. 121–143, 2008.

[4] M. Ajmal, M. Ashraf, M. Shakir, Y. Abbas, and F. Shah, "Video summarization: techniques and classification," In *Lecture Notes in Computer Science*, Vol. 7594, pp. 1–13, 2012

[5] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A Survey on Visual Content-Based Video Indexing and Retrieval," *IEEE T Syst.Man CY C*, vol.41, no.6, pp.797–819, 2011

[6] N. Dimitrova, H. J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of video content analysis and retrieval," *IEEE Multimedia*, vol. 9, no. 3, pp. 42–55, Sep. 2002.

[7] J. Iparraguirre, and C. Delrieux, "Speeded-Up Video Summarization Based on Local Features," In *IEEE International Symposium on Multimedia*, Anaheim, CA, USA, 2013, pp. 370-373.

[8] N. Ejaz, T. B. Tariq, and S. W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism," *J. Vis. Commun. Image Represent*, vol. 23, no. 7, pp. 1031-1040, 2012.

[9] G. H. Song, et al., "A novel video abstraction method based on fast clustering of the regions of interest in key frames," *AEU-Int. J Electron C*, vol. 68, no. 8, pp.783-794, 2014.

[10] S. E. F. de Avila, et al., "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognit. Lett.,* vol. 32, no. 1, pp. 56–68, 2011.

[11] S. K. Kuanar, et al., "Video key frame extraction through dynamic Delaunay clustering with a structural constraint," *J. Vis. Commun. Image Represent*, vol. 24, no. 7, pp. 1212-1227, 2013.

[12] K. M. Mahmoud, M. A. Ismail, and N. M. Ghanem, "VSCAN: An Enhanced Video Summarization Using Density-Based Spatial Clustering," *in Image Analysis and Processing Conference-ICIAP 2013*, vol. 1, pp. 733–742, 2013.

[13] J. Almeida, N. J. Leite, and R. S. Torres, "Online video summarization on compressed domain," *J. Vis. Commun. Image Represent.*, vol. 24, no. 6, pp. 729–738, 2013

[14] R. Goldman, R. Pea, B. Barron, and S. J. Derry (Eds.), "Video research in the learning sciences,", Mahwah, NJ: Lawrence Erlbaum, 2007

[15] S. Bekhet, A. Ahmed, and A. Hunter, "Video matching using DC-image and local features," *Lecture Notes in Engineering and Computer Science*, vol. 3, pp. 2209-2214, 2013.

[16] S. Li, J. J. Ahmad, D. Saupe, and C.-C. J. Juo, "An improved DC recovery method from AC coefficients of DCT-transformed images," in *Proc. IEEE Int. Conf. Image Process.*, Sept. 2010, pp. 2085–2088.

[17] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: a large dataset for non-parametric object and scene recognition," *IEEE T Pattern Anal*, vol. 30, no. 11, pp. 1958–1970, 2008.

[18] A. C. Bovik, Ed. "Handbook of Image and Video Processing," New York: Academic, 2000.

[19] J. Huang and Y. Q. Shi, "An adaptive image watermarking scheme based on visual masking," *Electron. Lett.*, vol. 34, no. 8, pp. 748–750, 1998.

[20] D. Androutsos, K. N. Plataniotis, and A. N. Venetsanopoulos, "Distance measures for color image retrieval," in *Proc. IEEE Conf. Image Processing*, vol. 2, Chicago, IL, Oct. 1998, pp. 770–774.

[21] I. Koprinska and S. Carrato, "Temporal video segmentation: a survey," *Signal Process.: Image Commun*., vol. 16, no. 5, pp. 477–500, Jan. 2001

[22] J. S. Boreczky and L. Rowe, "Comparison of video shot boundary detection techniques," in *Proc. IS&T/SPIE Storage and Retrieval for Still Image and Video Databases IV*, vol. 2670, Feb. 1996, pp. 170–179.

[23] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Syst*., vol. 1, pp. 10–28, 1993.

[24] X. Hou, J. Harel, and C. Koch, "Image Signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell*, Vol. 34, no. 1, pp. 194-201, Jan. 2012.

[25] N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Trans. Syst. Man Cybern*., vol. 9, pp. 62–66, Jan. 1979.

[26] Video Summarization Project, https://sites.google.com/site/vsummsite/. Accessed 24 June 2016

[27] Fast and effective video summarization, http://sites.google.com/site/fastvideosummarization/. Accessed 24 June 2016

[28] N. Ejaz, T. B. Tariq, and S. W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism," *J. Vis. Commun. Image Represent.*, vol. 23, no. 7, pp. 1031–1040, 2012.

[29] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pp. 226-231, 1996.

Antonio Cedillo Hernandez was born in Mexico. He received the B.S. degree in Computer Engineering, the M.S. degree in Microelectronic Engineering and his PhD in Communications and Electronic from the National Polytechnic Institute of Mexico in the years 2005, 2007 and 2013, respectively. He has about seven years of professional practice in several strategic positions related to IT. Currently, he has a postdoctoral position at National Autonomous University of Mexico. His principal research interests are video and image processing, information security, watermarking and related fields.

Manuel Cedillo Hernandez was born in Mexico. He received the B.S. degree in Computer Engineering, the M.S. degree in Microelectronics Engineering and his PhD in Communications and Electronic from the National Polytechnic Institute of Mexico (IPN) in the years 2003, 2006 and 2011, respectively. He has six years of professional experience at Government positions related to IT. From September 2011 to December 2015 he was with the Engineering Faculty of the UNAM where he was a Professor. Currently, he is a full-time researcher at IPN. His principal research interests are image and video processing, watermarking, software development and related fields.

Francisco Garcia Ugalde was born in Mexico. He received the B.S. degree in 1977, in electronics and electrical system engineering from UNAM, his Diplôme d'Ingénieur from SUPELEC France in 1980, and his PhD in 1982 in information processing from Université de Rennes I, France. Since 1983, he is a full-time professor at Engineering Faculty, UNAM. His current research interest fields are: Digital filter design tools, analysis and design of digital filters, image and video processing, theory and applications of error control coding, turbo coding, cryptography applications, watermarking, parallel processing and data bases.

Mariko Nakano Miyatake was born in Japan. She received the M.E. degree in 1985, in Electrical Engineering from the University of Electro-Communications, Tokyo Japan, and the PhD degree in Electrical Engineering from Metropolitan Autonomous University (UAM), Mexico City, in 1998. From July 1992 to February 1997 she was at Department of Electrical Engineering in UAM. In February 1997, she joined the Graduate Department of The Mechanical and Electrical Engineering School at National Polytechnic Institute of Mexico, where she is now a Professor. Her research interests are in information security, image processing, pattern recognition and related fields.

Hector Perez Meana was born in Mexico. He received his M.S: Degree on Electrical Engineering from the Electro-Communications University of Tokyo Japan in 1986 and his PhD degree in Electrical Engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1989. From March 1989 to September 1991, he was a visiting researcher at Fujitsu Laboratories Ltd, Kawasaki, Japan. From September 1991 to February 1997 he was with the Electrical Engineering Department of the UAM where he was a Professor. In February 1997, he joined the Graduate Studies and Research Section of The Mechanical and Electrical Engineering School, of the National Polytechnic Institute of Mexico, where he is now a Professor. His principal research interests are adaptive systems, image processing, pattern recognition, watermarking and related fields.