A SCALABLE APPROACH TO VIDEO SUMMARIZATION

Ayman Bassam Nassuora
Dept. of Management Information Systems City University College Ajman (CUCA) Ajman, United Arab Emirates
a.bassam@cuca.ae

ABSTRACT—*Algorithms for the generation of video summaries allow us to obtain synthesized sequence information contained herein. In this context, the duration of summaries plays a key role because sometimes synopsis with varying length is desired, so the scalable summaries allow us to respond to this need. In this paper, a new procedure is developed with the aim of obtaining scalable video summaries according to their duration. The procedure developed is based on a process of creating the summaries with well-established steps that guarantee efficient generation thereof. Results show compatibility with various encoding formats, effectively generating summaries for various lengths. The effectiveness of the procedure, confirmed by the results of the work ensures their potential application in systems for management, processing and transmission of audiovisual materials.*

Keywords: Scalable Video; Visual Summaries; Advanced Video Coding; MPEG

I. INTRODUCTION

In recent years there has been a marked increase in the production and distribution of audiovisual content. Studies show that internet users prefer and mostly consume audiovisual archives available online [1, 2] sites like YouTube, dedicated to sharing audiovisual materials in the network receive over two million visits on a daily basis. This, coupled with increased computing capabilities and storage systems, has set us the challenge of improving the environments of production, distribution and retrieval of audiovisual content, as it is predicted to have continuing rise in demand for this medium of content.

Access to multimedia content is an increasingly difficult activity which usually involves a process of pre-visualization of the video by the user. In this context, video processing techniques have become a necessity due to the sheer volume of audiovisual materials available. Several types of research have been devoted to establishing procedures to describe, process, store Access to multimedia content is an increasingly difficult activity which usually involves a process of pre-visualization of the video by the user. In this context, video processing techniques have become a necessity due to the sheer volume of audiovisual materials available. Several types of research have been devoted to establishing procedures to describe, process, store and retrieve information from audiovisual content. Methods for generating automatic summaries of video let us create a compacted visual representation of information in the video, i.e. provide a synthesized sequence, but with representations of original content [3, 4, 5, 6].

Video summaries are designed to facilitate navigation within a database of content. We can even use the summaries obtained as a final product that guarantees the user to quickly access relevant semantic positions in the sequence [4, 7]. Most of the algorithms developed in this area generate a single output sum. This may be insufficient because sometimes a personalized representation for each user is desired. Generating scalable summaries can be a useful way to address the diversity of preferences and to achieve greater usability.

In this paper, scalability is assumed as the ability to produce multiple output summaries with different durations for the same original sequence. As a background to this research, we must first observe the approaches aimed at achieving summaries of various scales [3, 8, 3, 9].

The approaches mentioned above have shown good results but have had certain difficulties. [9] assumed scalability for each hierarchical structure with the limitation that the summaries are not scalable within each level. For the results presented by [3] the algorithm is based on MPEG standards [10]. This procedure focuses on the characteristics of the data stream of MPEG coding, which limits their use in other encoding schemes. Despite the difficulties of the studies analyzed, the concepts of scalability and the embedded sum raised in this approach [3] are the most widely accepted for the creation of scalable video summaries.

This research has its origins in the applications for managing, processing and transmission of audiovisual content. These applications are characterized by the steady increase in the number and diversity of managed and subsequently stored for use audiovisual materials. Audiovisual materials are not necessarily managed code in a prescribed format as this depends on the requirements of potential customers.

Moreover, in management applications, processing and transmission of audiovisual content should ensure the cataloging process of audiovisual materials managed for easy search and proper use. Algorithms have been developed to achieve semi-automatic cataloging, but often continue manual annotation of certain information still needed. In this process, users gain the perspectives of audiovisual content browsing, sometimes randomly on video sequences. In addition to these applications in order to provide users of audiovisual materials available on the web to consume on-demand pursued. It is common to notice in users consuming video, usually of long duration, prior to the full display of the material, perform a preview of it to get an idea of its contents and then visualize it completely if they consider it to your liking.

The foregoing reveals the need for a mechanism for several synthesized sequences with significant information from the same video, favoring the process of cataloging and pre-visualization of audiovisual materials management systems, processing and transmission of audiovisual files. In this paper, a procedure developed to generate scalable video summaries exposed.

II. METHODOLOGY

Creating video summaries involves various levels of complexity, in this paper the results of the study [6], where the sequence of steps required evidence for a video summary assumed. In the following diagram the sequence of computational steps for transforming a sequence of video input, a summary of it is displayed.

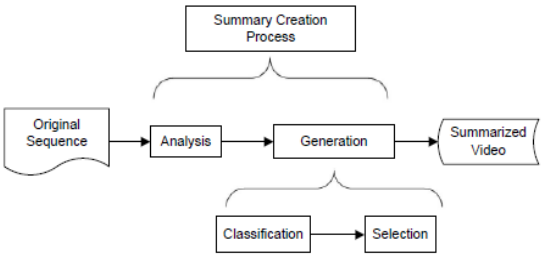

Fig.1 Steps to create a video summary

III. ANALYSIS STAGE

In the analysis phase operations to extract visual descriptors of the original sequence in order to obtain a list of the data and the descriptors associated with the shots are taken. At this stage the video is decoded and frames thereof are extracted. Once captured frames proceed to the first phase to segment video, shot detection histograms [11, 12]. This is divided into four steps:

1) Calculate the histogram of each frame.
2) Compare each frame with the adjacent and save the results.
3) Establish a threshold between comparisons.
4) Detect all comparisons that exceed this threshold as a tap change.

At this point, the video has been segmented using color histograms but this procedure of shot detection is very sensitive to changes in illumination. Therefore the detection is performed by edges to frames that represent the change from one shot to another [13, 11, 14]. This second phase of detection is achieved by implementing the following steps:

5) Convert the frames to grayscale
6) Perform a Gaussian blur.
7) Apply a function of edge detection.
8) Expand the resulting lines.
9) Compare the results.
10) Establish threshold comparisons.
11) Detect false positives from the previous step.

After making detection by histograms, edges can still be cut within a shot that was not detected. This is the reason why the SURF descriptor is used for those limits between takes that as a result of the analysis of histograms belong to a dubious margin [15, 7, 14, 16, 13]. The implementation of this third phase detection is achieved by the following steps:

1) Remove the features.
2) Get the descriptors.
3) Search matches between frames.
4) Thresholding the result.
5) False positives from the previous step (Detect).

Finally, as a result of this stage a list of data for decision is obtained:

- Starting position.
- Length.
- Keyframe.
- Threshold Histogram color.
- Threshold SURF

The analysis stage is the critical step in the whole process because the resulting list is stored and constitutes the input to the generation stage. Further analysis will not run again because although it is required to generate another summary this is guaranteed with the stored data. The analysis is generated once and as many times as necessary, i.e. only be performed the first time a summary is required, scalability is ensured after running it several times. Next-generation is explained in detail.

IV. SCALABLE VIDEO GENERATION

The step of generating the higher complexity lies in the creation of a video summary. For this reason, this stage is divided into two: classification and selection [6]. During classification provides a grade on each frame, depending on the features discussed above. In the selection, step determines which of the shots, be included in the summary of video to generate, and which you may be waived by classification made previously.

At work scalability criteria set out in assumed [3] which defines a scalable summary is that which is constituted by a group of embedded summaries, $SS=\{S_l,S_{l+1}, \ldots, S_{L-1}, S_L\}$ where $l \varepsilon N$ denotes the scale summary and L the length thereof. In turn, compliance with the following restriction is

necessary $S_l \subset S_{l+1} \subset S_{L-1} \subset S_L$, Whereas each summary smaller scale is embedded in the summary of a larger scale. As you can see the summary of greater length is formed by grouping summaries shorter, so it must achieve a hierarchy in which the teaser shorter ensure greater representation of the original sequence.

At this stage the aim is to obtain the most representative shots of the video, trying to eliminate redundancy by providing as much information as possible. With the results of the analysis of color histogram and SURF proximity matrix [7] is constructed, this matrix is used later to create a group of K cluster $\{C_o,C_1, \ldots, S_{K-1}\}$ according to the distances between each pair of jacks $d(t_i,t_j)$ grouping shots with similar features. Then for each cluster C_k closest to the centroid shots, determined by a score established according to duration and variability in the group, are considered the most representative group, which provides them greater priority to include in the summary, settling the necessary hierarchy for scalability [17, 18].

For the final generation summarizes the two modes for displaying video summaries that can be commonly observed are assumed:

1) Summaries built with static images, which are a representation of the relevant frames, taken from the original video sequence to display the contents thereof through a static storyboard, commonly known as a storyboard. This mode is defined in [4] as follows: $R=A_{KeyFrame}(V)=\{f_1, f_2, \ldots, f_k)$ where $A_{KeyFrame}$ is the method of extracting the video V, keyframes, yielding a representation R constituted by frames $\{f_1, f_2, \ldots, f_k)$
2) Abstracts which consist of short video segments. This summary mode is a set of video segments are extracted from the original video, which is grouped either by a court or through a gradual transition effect for a video smaller than the original. In this case, a small video called skimming video is constituted. The video skimming defined in [4] as follows:

$K=A_{Skim}(V)=E_{i1} \cup E_{i2} .. \cup E_{ik}$

where A_{skim} is the method of generation of a skimmed version of the video V, $E_i \varepsilon V$ is the

ith fragment to include in the video skim and ∪ obtaining the integration operation a representation K, constituted by the integration of the segments

E_{i1}, E_{i2}, …, E_{ik}. This integration is performed following the temporal flow of the original video and the integration operation used is a transition from simple disappearance. This step is performed to obtain the desired length summaries, therefore, can be repeated as often as necessary, to ensure conditions of scalability.

V. RESULTS AND DISCUSSION

In order to test the proposed procedure proceeds to try on videos with different characteristics, so that non-probabilistic intentional sample videos compiled by the author videos selecting different encodings used. This test is done to validate the compatibility of the procedure developed with video encoded in different formats. Among the features of these videos are the container, the encoder, the bitrate1, frames per second (FPS) and resolution.

The result of testing format support for performing the method with four videos with different encodings shown

Table I. Test For Format 1

Name of Video	Arsenal Highlights
Container Format	AVI
Encoder	MPEG4-Xvid
Bitrate (Kbps)	128
Frames per second	29.97
Resolution	800 x 600
Remarks: the video is processed correctly	

Table Ii. Test For Format 2

Name of Video	Chelsea Highlights
Container Format	WMV
Encoder	WMV2
Bitrate (Kbps)	2400
Frames per second	30
Resolution	720 x 480
Remarks: the video is processed correctly	

Table Iii. Test For Format 3

Name of Video	News Sequence
Container Format	MPG
Encoder	MPEG1
Bitrate (Kbps)	1150
Frames per second	25
Resolution	720 x 480
Remarks: the video is processed correctly	

Table iv. Test For Format 4

Name of Video	Pottery Tutorial
Container Format	ogg
Encoder	Theora
Bitrate (Kbps)	1596
Frames per second	29
Resolution	352 x 240
Remarks: the video is processed correctly	

Table v. Test For Format 5

Name of Video	Coldplay Song
Container Format	webm
Encoder	Vp8
Bitrate (Kbps)	1596
Frames per second	29
Resolution	352 x 288
Remarks: the video is processed correctly	

As shown (see Tables 1-4) satisfactory to generate video summaries with different encodings results are obtained, this result shows greater efficacy with respect to work done by [3] it shows that the developed procedure can use in videos with coding that is not based on the MPEG standards.

In the analysis stage obtaining all necessary data for the generation of short, is for this reason that validates this stage is determined independently by the process of segmentation is performed on the same guarantees, constitutes a section review in the proposed procedure. Then developed tests are performed in order to validate the shot detection during analysis using recall precision measures. For this analysis, the video database created by Bescos is used.

Table VII

	No. of Frames	Cuts	Cuts Detected	Fp	Fn	Recall	Precision
V1	30715	54	51	4	3	94%	93%
V2	27643	13	12	1	0	100%	92%
V 3	90341	21	21	2	0	100%	91%
V 4	114997	127	125	6	2	98%	96%

As shown (see Table 5) in the analysis stage of the procedure developed for shot detection Precision guaranteed rates above 90% in all cases, which can be considered for these procedures accepted [19, 20].

Scalable Generation Tests

Below is a table showing the results of the proposed tests, for generating synthesized several video sequences the same procedure, taking into account the variation of the length of said sequences. Various scales for both storyboard generation to generation video skimming set. For this analysis, we have established the following levels: 20, 30 and 40 number of frames of the original sequence, for generating a storyboard and 10, 20 and 30 percent of the original sequence for the case of skimming video. Table time in seconds it takes for the procedure for generating the summary for established scales is. **Table VII. Test for the generation of scalable summaries**

	Generation					
	Storyboard (#)			**Skiming (%)**		
V1	20	30	40	10	20	30
	0.67(s)	0.74(s)	0.81(s)	0.98(s)	1.36(s)	1.64(s)
V2	20	30	40	10	20	30
	0.68(s)	0.76(s)	0.79(s)	1.02(s)	1.46(s)	1.87(s)
V3	20	30	40	10	20	30
	0.67(s)	0.72(s)	0.79(s)	0.99(s)	1.33(s)	1.63(s)
V4	20	30	40	10	20	30
	0.69(s)	0.73(s)	0.82(s)	0.97(s)	1.29(s)	1.61(s)

As in the analysis stage, for testing the generation of scalable video database created by Bescos was used. The results show that the proposed method offers the possibility to create summaries of various scales, ie obtain adequate summaries with different lengths and acceptable response times. Note that the required number of doubles scalability and not necessarily in correspondence doubles the time required; this is because the analysis step to a larger scale is not executed. Even obtaining these results should not be ruled out other visual descriptors used to improve process efficiency achieving superior results.

Validation of Usability of Summaries

To ensure the usability of an abstract video, you must create a teaser video that has a visual language that meets two basic conditions: semantic coverage and visual pleasure. The first condition relates to the created summary to preserve as much information as possible representatively, discarding as much redundant information, in order to minimize the time required for display. The second of these conditions means that the abstract should not only be informative but must have nice visual features to the user who finally displayed [21, 4].

To validate the usability of the summary are proven ability to provide the user with a comprehensive understanding of the information contained in the original video, ability to provide a synthesis of information from the original video by the representative contents prevail on the verbal and the ability to achieve a chain mosaic in which shots or images that compose the summary together, make sense.

To validate the performance of these two conditions survey was applied as a diagnostic tool. A technique of simple random sampling [16] considering a population consisting of 90 users SIAV is selected. To determine the sample size used the software Stats considering a confidence level of 95%, settling to apply the survey to 30 users. The survey consists of four questions applied where the user can assign a real number between 1 and 10 on establishing its usability final summary rating.

For analysis of the survey results is processed by averaging the values obtained in each of the 30 users' questions by obtaining the results shown in the following table.

Table viii. Average Values Obtained from the Survey

Questions	1	2	3	4
Avg. Values	6.1	5.3	4.7	6.6

Averaging the values obtained after a level of usability of 5.67 is obtained. To determine the level of usability of a scale summarized in numerical values and nominal equivalent shown in the table

Table ix. Scale for Level of Usability

Numerical	0-1.9	2-3.9	**4.5-9**	6-7.9	8-10
Nominal	Very Low	Low	**Avg**	High	Very High

The results show that the procedure meets moderately with the agreed terms of usability. It should continue to refine it because, in the ability to provide semantic coverage and visual pleasure to the end-user, the procedure results show that inefficiencies persist. As a consequence of this result, it is recommended to redefine the step of generating the proposed procedure.

VI. CONCLUSION

The proposed procedure allows efficient generation of scalable video summaries addressing the length of the sequences synthesized as scalability requirements. It also supports the summarization of videos that are not encoded in MPEG standards.

The steps involved in the generation of an abstract set properly with good results in the analysis stage, where the detection of changes of sample shots and recall Precission rates above 90% in all cases. Also at this stage to obtain the data necessary for the generation scalable without analyzing the original video again is guaranteed. For his part in the step of generating summaries obtaining different lengths is achieved by ensuring scalability and achieving the desired representation abstracts patterns skimming storyboard and video.

The main difficulties of the proposed method are demonstrated in the ability to adequately achieve semantic coverage and visual pleasure for the end-user, which implies the need to redefine the stage set generation.

It is recommended to consider working with other visual descriptors which can improve the results. Although the results show the need to continue working on the proposed method, it is considered that can be used in systems management, processing and transmission of audiovisual content, minimizing the limitations of this system for several sequences synthesized with significant information the same video which will benefit the process of cataloging and pre-visualization of audiovisual materials.

VII. REFERENCES

[1] López-vidales, González-Aldea. and Medina-de-la-viña, "Jóvenes y televisión en 2010," *Un cambio de hábito*, vol. 30, no. 16, pp. 97-113, 2011.

[2] M. y. Maass and J. González, "De memorias y tecnologías radio, televisión e Internet en México," *Estudios sobrelas culturas contemporáneas*, vol. 22, pp. 193-220, 2005.

[3] J. L. Herranz and M. Martínez, "A framework for scalable summarization of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 9, pp. 1265-1270, 2010.

[4] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM*, vol. 3, no. 1, pp. 1-37, 2017.

[5] O. Paul, A. F. Smeaton and G. Awad, "The trecvid 2008 BBC rushes summarization evaluation," in *TVS '08 Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, Vancouver, British Columbia, Canada, 2008.

6] V. Valdés and J. M. M. Sanchez, "On Video Abstraction Systems' Architectures and Modelling," in *Semantic Multimedia, Third International Conference on Semantic and Digital Media Technologies, SAMT 2008*, Koblenz, Germany, 2008.

[7] P. Berkhin, "A survey of clustering data mining techniques: grouping multidimensional data," *Springer*, pp. 25-71, 2006.

[8] S. Benini, A. Bianchetti, R. Leonardi and P. Migliorati, "Extraction of significant video summaries by dendrogram analysis," in *Internation Conference on Image Processing*, Atlanta, GA, 2006.

[9] X. Zhu, "Exploring video content structure for hierarchical summarization Multimedia Systems,," *Multimedia Systems*, vol. 10, pp. 98-115, 2004.

[10] Mitchell and J. L., "MPEG Video Compression Standard," Kluwer Academic Publishers, Holladay, UT, USA, 1996.

[11] C. Yinzi, D. Yang, G. Yonglei, W. Wendong, Z. Yanming and W. Kongqiao, "A Temporal Video Segmentation and Summary Generation Method Based on Shots' Abrupt and Gradual Transition Boundary Detecting," *Communication Software and Network*, 2010.

[12] D. A. Díaz Espinoza, Implementación y comparación de descriptores para búsquedas en video.: Facultad de Ciencias Físicas y Matemáticas. Departamento de Ciencias de la Computación., de Chile, Universidad de Chile, 2011.

[13] R. Lienhart, "Reliable transition detection in videos: A survey and practitioner's guide," *International Journal of Image and Graphics*, vol. 1, no. 3, pp. 469-486, 2001.

[14] J. e. a. Bescós, "A unified model for techniques on video-shot transition detection," *IEEE Transactions on Multimedia*, vol. 7, no. 2, 2005.

[15] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, p. 177–280, 2008.

[16] H. Bay, Tuytelaars and Van, "SURF: Speeded up robust features," in *Computer Vision - ECCV 2006*, Graz, Austria, 2006.

[17] E. Dumont and B. Mérialdo, "Redundancy removing by adaptive acceleration and event clustering for video summarization," in *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, Klagenfurt, Austria, 2008.

[18] R. Xu and D. W., Clustering, Wiley-IEEE Press, 2009.

[19] P. Mohanta, S. Saha and B. Chanda, "A heuristic algorithm for video scene detection using shot cluster sequence analysis," in *Seventh Indian Conference on Computer Vision, Graphics and Image Processing*, Chennai, India, 2010.

[20] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *23rd international conference on Machine learning (ICML)*, Pittsburgh, Pennsylvania, 2006.

[21] J. Ren, J. Jiang and Y. Feng, "Activity-driven content adaptation for effective video summarization," *Journal of Visual Communication and Image Representation*, pp. 930-938, 2010.

[22] J. Bescós, G. Cisneros, J. M. Martínez, J. M. Menéndez and Julián, "A unified model for techniques on video-shot transition detection," *IEEE Transactions on Multimedia*, vol. 7, no. 2, pp. 293-307, 2005.

[23] V. Valdes and M. José, "On Video Abstraction Systems' Architectures and Modelling," in *International Workshop on Visual Content Processing and Representation*, Berlin, 2008.

[24] Herranz Arribas, "A scalable approach to video summarization and adaptation," *Escuela Politécnica Superior de Ingeniería Informática. Madrid, Universidad Autónoma de Madrid*, p. 224, 2010.