Multiscale Non-Linear Video Content Decomposition: Analysis and Evaluation

Anastasios D. Doulamis and Nikolaos D. Doulamis
National Technical University of Athens, Department of Electrical and Computer Engineering, E-mail: adoulam@cs.ntua.gr

ABSTRACT

In this paper, we theoretically analyze the efficiency of a multiscale non-linear video decomposition scheme, which is used for progressive retrieval and navigation of video sequences (hierarchical summarization). In particular, the efficiency is expressed as the difficulty for a user to locate a relevant video segments, while moving through different levels of hierarchy. It is proved that multiscale video organization is more efficient than the sequential one, even in the worst case where the organization is not performed on a visual content perception basis. However, much greater improvement is achieved in case that video organization is accomplished with respect to the human perception of the visual content based on visual content classification algorithms. Experimental evaluation of multiscale organization is presented with respect to sequential scanning for different visual content organization and video decomposition methods.

1. INTRODUCTION

Traditionally, video information is represented as a sequence of consecutive frames, each of which corresponds to a constant time interval. While such a linear (sequential) representation is adequate to "play" a video in a movie mode, it is not appropriate for interactive navigation of video information over networks. Currently, the only way to exchange visual data among different databases, or to remotely access video archives located on distributed platforms, is to perform either video streaming or video file downloading, techniques which are both tedious and time consuming.

To overcome the aforementioned difficulties progressive schemes for video decomposition should be introduced so that the visual information is transmitted at different "resolution levels" in a hierarchical framework. The idea stems from the method used for progressive transmission of still images. In this case, instead of transmitting the image sequentially at a full resolution, by scanning it line by line, a lower image resolution is first delivered and then, the image quality gradually enhances so that the user is able at any time to see a preview of the image content [1]. Hierarchical video encoding has been recently adopted in the MPEG-4 standard to increase the flexibility of video streaming based on a Fine Granularity Scalability (FGS) scheme [1][2]. This concept can be extended to video sequences by initially transmitting a coarse (low) resolution and then, decomposing particular video segments into a higher (finer) resolution. In this framework, the user can interweave in the process to determine segments of interest that can be further decomposed, resulting in an interactive non-linear navigation of video sequences.

Algorithms for video summarization can be considered as the first attempts towards a non-linear video representation [3]-[7]. However, video summarization or abstraction schemes are not appropriate for interactive video navigation over networks [8]. This is due to the fact that the goal of these algorithms is to extract a small "video abstract" by discarding visual information, similarly to text summaries used in document files. Suppose, for example, a physician who accesses remote medical archives of video sequences in order to find a set of frames, e.g., a shot, which he/she is interested in. In this case, application of a video summarization scheme may discard the frames of physician's interest. On the contrary, in a video navigation scheme, no information should be lost since any video segment can be considered to be of particular users' interest.

Towards this direction, algorithms, dealing with progressive retrieval of images or video sequences, such as the ones presented in [8] and [9], can be considered. These techniques initially transmit a coarse (low) "resolution view" of an image, followed by additional residual information. In particular, in [9], images are decomposed in space and frequency domain, while in [8], extension to video sequences has been investigated using spatial-temporal filter banks. However, these approaches, linearly decompose (downsample) visual information on spatial and temporal direction *at fixed units* of pixels or frames. Thus, video is not organized according to the visual content information. Hierarchical summarization has been adopted in the framework of the MPEG-7 standard [10], through the HierarchicalSummary Description scheme. The standard provides a syntax for describing a hierarchical video organization and suggests an algorithm for constructing the hierarchical summaries. The technique is based only on a key-frame organization and clusters video segments according to the visual content and temporal coherency [10].

A more efficient video decomposition can be accomplished if video is hierarchically analyzed in the *"content" domain*, meaning that visual information is partitioned into segments of relevant content and thus different content resolution levels are created [11] In particular, video decomposition is represented as a tree structure, the levels of which indicate the respective content resolution, while the nodes correspond to the segments that video is partitioned at this level. In this framework, the user is able to select segments (tree-nodes) of interest and reject segments of non-interest. For all selected segments, further visual decomposition is accomplished, resulting in a content hierarchy from the lowest (coarse) to the highest (fine) content resolution.

In this paper, we theoretically analyze the efficiency of a multiscale video decomposition scheme. In particular, the efficiency of a non-linear multiscale video decomposition is measured as the "difficulty" for a user to locate relevant video segments, while moving through different levels of the hierarchy. It can be proved that in case of a 4-level multiscale non-linear video organization, the efficiency is, *in the worst case*, 8 times greater than the traditional sequential video organization. This is, however, a lower bound of improvement and it is valid only in case that the visual content is randomly organized. Therefore, a *much greater improvement* is expected in case that video organization is accomplished with respect to the human perception of the visual content. Since the main issue that affect the efficiency is the number of paths required for a user to locate a relevant video segment, better visual content organization increase the efficiency of a multiscale video organization scheme. This is presented in this paper, where an experimental evaluation of the improvement ratio of a multiscale organization and a sequential video scanning is presented.

2. VIDEO CONTENT DECOMPOSITION

A set of representative shots is initially extracted from a video sequence to form a coarse (low) resolution of its visual content. Then, the remaining shots are classified with respect to these shot-

representatives, creating shot-classes. Let us denote as γ_i i=1,2…,P_s, the P_s shot representatives of a sequence V and as S_i, i=1,2,…,P_s, the respective shot classes. Index i corresponds to the shot class index and not to the shot sequence index. Since each shot is assigned only to one class, it is held that

$$S_i \cap S_j = \varnothing \text{ with } i \neq j \text{ and } V = \bigcup_{i=1}^{P_s} S_i \quad (1)$$

Let us denote as $s_{l,k}$ the lth shot of class S_k. Thus,

$$S_k = \bigcup_{l=1}^{M(S_k)} s_{l,k} \quad (2)$$

where $M(S_k)$ corresponds to the number of shots of the class S_k. It should be mentioned that, in our notation, index l of $s_{l,k}$ refers to the lth shot (element) of class S_k and not to the lth shot of video sequence V since class S_k does not contain temporal consecutive shots. The P_s shot representatives as well as the respective shot classes S_k , k=1,2,…, P_s are estimated as described in section 4.

Similarly, the content of a shot can be further decomposed into frame-classes by extracting a set of representative frames from this shot and then classifying all the remaining frames with respect to these representatives. Let us denote as $\delta_m^{(l,k)}$, m=1,2,…, $P_f^{(l,k)}$ the $P_f^{(l,k)}$ representative frames of the shot $s_{l,k}$ and as $F_m^{(l,k)}$, m=1,2,.., $P_f^{(l,k)}$ the corresponding frame classes of this shot, which satisfy the following equation,

$$F_m^{(l,k)} \cap F_n^{(l,k)} = \varnothing \text{ with } m \neq n \text{ and } s_{l,k} = \bigcup_{m=1}^{P_f^{(l,k)}} F_m^{(l,k)} \quad (3)$$

Equation (3) indicates that each frame of shot $s_{l,k}$ is assigned only to one frame-class. In a similar way, we denote as $f_{i,m}^{(l,k)}$ the ith frame of the class $F_m^{(l,k)}$. Thus,

$$F_m^{(l,k)} = \bigcup_{i=1}^{N(F_m^{(l,k)})} f_{i,m}^{(l,k)} \quad (4)$$

where $N(F_m^{(l,k)})$ returns the number of frames belonging to class $F_m^{(l,k)}$. The $P_f^{(l,k)}$ frame representatives and the respective frame class $F_m^{(l,k)}$ are estimated as described in section 4.

In this way, a video sequence is hierarchically decomposed into different content-resolution levels so that any video component, i.e., frame or shot, can be accessed, in a non-sequential manner, according to its relevant content in the sequence. For example, to access the frame $f_{i,m}^{(l,k)}$, the user should sequentially select the shot-representative γ_k and thus the shot-class S_k, followed by the shot $s_{l,k}$ of this class and finally followed by the frame class $F_m^{(l,k)}$ of shot $s_{l,k}$. Figure 1 presents a simple example of the proposed progressive content-based video decomposition scheme using the proposed tree structure hierarchy.

3. QUALITY EVALUATION

In this section, we analyze the efficiency of a multiscale content organization as the "difficulty" for a user to indeed find what he/she is looking for, while moving through different levels of the hierarchy.

Let us first consider on the traditional (sequential) video organization. In this case, the information required to be transmitted for accessing the ith frame of a video sequence V is $i \cdot p_i$, where p_i refers to the probability of considering the ith video frame as frame of interest. Assuming that all frames present the same probability to be selected, we have that $p_i = 1/P$, where P denotes the number of frames of the video sequence V. Then, the average difficulty is given by

$$E_l = \sum_{i=1}^{P} i \cdot p_i = \frac{P+1}{2} \approx \frac{P}{2} \quad (5)$$

since $\sum_{i=1}^{P} i = P \cdot (P+1)/2$ and usually $P >> 1$. Equation (5) means that, in the linear (sequential) case, half of the video frames should be on average transmitted to access a relevant frame.

Let us now, for simplicity, we consider a two-level video hierarchy. This means that the video sequence V is decomposed into frame-representatives and then upon a user's selection, all frames belonging to the respective frame class are sequentially (linearly) transmitted. Let us denote as $\tilde{N}$ the average number of frames belonging to the frame classes, i.e., $\tilde{N} = E\{P(G_m^{(l,k)})\}$, where $E\{\cdot\}$ is the expectation operator. Then, the difficulty (efficiency) $E_m^{(2)}$ is given by

$$E_m^{(2)} = K_f + \frac{(\tilde{N}+1)}{2}(1 + c_2 + c_3 + \cdots + c_{K_f}) \quad (6)$$

where K_f is number of frame representatives. The c_i indicates the probability that the frame of user's interest *does not* belong to the class of the 1st and the 2nd, …,and the (i-1)th selected representative frame. Equation (6) indicates that initially all frame representatives are transmitted. Then, for each selected representative, all frames of the respective class are transmitted in a linear way, meaning that the difficulty is on average ($\tilde{N}$ +1)/2. This is, however, valid only in case that the *first selected* class contains the frame of users' interest. Otherwise, another selection is made and the frames of the second selected class are linearly transmitted until the frame of interest is found. The efficiency of the second selection is multiplied by c_2, which expresses that probability of not finding the frame of interest in the first selection and so on.

As can be seen from Equation (6), the difficulty $E_m^{(2)}$ depends on probabilities c_i. The best case results from $c_2, c_3, ..., c_{K_f} = 0$. This means that only one selection is adequate for localizing the frame of interest. On the contrary, the worst case, is obtained when a frame of interest presents the same probability of belonging to any frame class. Thus, in the worst case, it is held that

$$c_i = 1 - (i-1)/K_f \quad (7)$$

Using (7), the efficiency $E_{m,wc}^{(2)}$ in the worst case equals

$$E_{m,wc}^{(2)} = K_f + \frac{(\tilde{N}+1)}{4} * (K_f + 1) \approx \frac{P}{4} \quad (8)$$

since $\tilde{N} * K_f = P$ and $P >> 1 + K_f + \tilde{N}$. Equation (8) means that, apart from a small overhead $K_f << P$, *in the worst case*, the

quarter of the total number of frames should be transmitted on average, instead of the conventional linear case, where the half of the total number of frames are required.

A greater reduction of the average transmitted information is achieved using the 4-level video decomposition scheme. In particular, the difficulty $E_m^{(4)}$ is given by

$$E_m^{(4)} = K_s + \tilde{M}\cdot(1+r_2+\cdots r_{K_s}) + \tilde{K}_f\cdot(1+r_2+\cdots r_{K_s})\cdot(1+b_1+\cdots+b_{\tilde{M}}) + \frac{(\tilde{N}+1)}{2}\cdot(1+c_2+\cdots+c_{\tilde{K}_f})\cdot(1+b_1+\cdots+b_{\tilde{M}})\cdot(1+r_2+\cdots r_{K_s}) \quad (9)$$

where $\tilde{M}$ refers to the average number of shots within a shot class, the $\tilde{K}_f$ to the average number of frame representatives within a shot and $\tilde{N}$ to the average number of frames within frame classes. We recall that K_s is the number of representative shots. The c_i, i=2,3…, $\tilde{K}_f$, corresponds to the probability that the frame of interest does not belong to the 1st and the 2nd, …, and the (i-1)th selected frame class of a shot. Similarity, the b_i, $i = 2,3,...,\tilde{M}$ refers to the probability that the frame of interest does not belong to the 1st and the 2nd,…, and the (i-1)th selected shot of a particular shot class, while the r_i i=2,3…,K_s, to the 1st, and the 2nd, …, and the (i-1)th selected shot class, defined by the respective shot representative. The best case results if c_i, b_i, r_i=0, $\forall i = 2,3,...$. Instead, the worst case results from the fact that a frame of interest has the same probability of belonging to any shot/frame class. Then,

$$c_i = 1-(i-1)/\tilde{K}_f,\ c_i = 1-(i-1)/\tilde{M} \text{ and } r_i = 1-(i-1)/K_s \quad (10)$$

Based on (10), the efficiency in the worst case is given as

$$E_{m,wc}^{(4)} = K_s + \tilde{M}\cdot\frac{(K_s+1)}{2} + \tilde{K}_f\cdot\frac{(\tilde{M}+1)\cdot(K_s+1)}{4} + \frac{(\tilde{M}+1)\cdot(K_s+1)\cdot(\tilde{N}+1)\cdot(\tilde{K}_f+1)}{16} \approx \frac{P}{16} \quad (11)$$

since the fourth term of equation (11) is much greater than the other terms. Thus, in the proposed 4-level structure video decomposition scheme, the average transmitted information, *in the worst case,* equals 1/16 of the total number of frames of the sequence, apart from an additional overhead.

The average transmitted information obtained from equation (11) is valid only in the worst case, and thus it indicates a bound of the average efficiency. This is due to the fact that the assumption that a frame of interest may belong to any frame/shot class with the same probability, is valid only in case that the representative frames/shots are randomly selected. The best performance is achieved if the probabilities c_i, b_i, r_i are distributed so that the greatest proportion is accumulated to the first selection. On the contrary, the probability that a frame of interest belong to the last selection should be almost zero. This is accomplished if video is organized based on visual content properties. Such a video organization maximize the probability of accessing a frame of interest in the first selection and therefore, the average transmitted information is as mush as possible near to the *best case* instead of the *worst case.*

The improvement ratio of the efficiency obtained using a multiscale video organization scheme over the sequential organization is

$$R = \frac{E_l}{E_m} \quad (12)$$

4. EVALUATION

In this section, we evaluate the performance of a multiscale video decomposition against several algorithms used for non-linear video content organization and representation. For the evaluation, we use three different MPEG coded video sources each of duration of 45 minutes. The sequences have been selected to represent different thematic areas.

Three different classification schemes are evaluated in this paper. In the first approach, initially content representatives are estimated (either key-shots or key frames). The content representatives are optimally extracted with a miximal discrimination based on a Genetic Algorithm (GA) scheme [11]. Then, the remaining shots/frames are classified with respect to these representatives. The second approach is based on a k-means organization. In this case, the number of classes are optimally estimated based on a maximum class separation method [5]. Finally, the method of [7] where grouping into story units are performed is investigated. Table I presents the results, where we observe that the method of [11] provides the best classification accuracy. Finally, the effect of the classification methods on the improvement ration R is presented in Table I. The R has been measured as the average over 3,000 different experiments for each of which the user searches for a frame of interest going through the different levels of the hierarchy or using the sequential scanning. In this experiment a 4-level multiscale video content decomposition is presented. Finally, in Table I presents the average R over 3,000 different experiments for several video decomposition/summarization schemes. decomposition/summarization methods. In all cases, the proposed 4-level video hierarchy outperforms the compared ones

Table II presents the improvement ratio for the traditional video summarization schemes. Particularly, in the works of [7], [5] R is significantly reduced compared to the proposed video hierarchy. This is due to the fact that these approaches estimate only frame representatives, and then clusters the remaining frames with respect to these representatives. Therefore, they ignore shot clustering resulting in a less efficient video decomposition. On the other hand, the hierarchical video decomposition of the work of [8] is based on a linear spatio-temporal video organization, which is less efficient than the proposed video content decomposition scheme. In addition, the suggestive algorithm of the MPEG-7 hierarchical summarization scheme [10] does not organize the video sequence in an optimal content oriented way, reducing the classification accuracy thus the decomposition efficiency.

In the following, we evaluate the efficiency of a 2 and 4-level video hierarchy, with respect to the distribution of the probabilities, c_i, r_i, b_i. Particularly, we assume a sequence of 40,000 frames, consisting of 400 shots, with an average number of key frames $\tilde{K}_f = 3$. Thus, $K_f = 1200$ and $\tilde{N} = 33$. We further assume that $K_s = 10$ and therefore $\tilde{M} = 40$. In this scenario, we modeled the probabilities c_i, r_i, b_i using the exponential density function, depending on a parameter β. As β increases, the classification accuracy decreases, and the probabilities tend to the form of equation (10). Instead, low values of β indicate high classification accuracy where only few number of paths are

required to find a frame of interest. Figure 2 presents the improvement ratio R with respect to parameter β for a 2 and 4-level content hierarchy. As can be seen, for high values of β the R tends to the worst case. On the contrary, low values of β yields high improvement ratios (better visual content organization). From Figure 2, it can also seen that a 4-level video hierarchy presents higher improvement ratios than a 2-level one.

5. REFERENCES

[1] W. B. Pennebaker and J. L. Mitchell, *JPEG: Still Image Data Compression Standard*. Van Nostrand Reinhold, New York, 1993.

[2] W. Li, "Overview of Fine Granularity Scalability in MPEG-4 Video Standard," *IEEE Trans. CSVT*, vol. 11, pp. 301–317, Mar. 2001.

[3] J. Nam and A. H. Tewfik, "Video Abstract of Video," Proc. of *the IEEE Inter. Workshop on Multimedia Signal Processing*, pp. 117-122, Copenhagen, Denmark, Sept. 2000.

[4] F. Arman, R. Depommier, A. Hsu and M.Y Chiu, "Content-based Browsing of Video Sequences", *ACM Multimedia*, pp. 77-103, Aug. 1994.

[5] A. Hanjalic and H. Zhang, "An Integrated Scheme for Automated Abstraction based on Unsupervised Cluster-validity Analysis," *IEEE Trans. on CSVT,* Vol. 9, No. 8, pp. 1280-1289, December 1999.

[6] N. Vasconcelos and A. Lippman, "A Spatiotemporal Motion Model for Video Summarization," *Proc. of IEEE CVPR*, pp. 361- 366, Santa Barbara, CA, June 1998.

[7] M. M. Yeung and B.-L. Yeo, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 7, No. 5, pp. 771- 785, October 1997.

[8] J. R. Smith, "VideoZoom: Spatio-temporal video browser," *IEEE Trans. on Multimedia,* vol. 1, No. 2, pp. 157-171, June 1999.

[9] J. R. Smith, V. Castelli and C.-S. Li, "Adaptive storage and retrieval of large compressed images," in *Storage & Retrieval for Image and Video Databases, VII*, M.M Yeung, B.L. Yeo and C. A. Bouman Eds. *Proc. SPIE*, vol. 3656, pp. 467-487, Jan. 1999.

[10] ISO/IEC JTC 1/SC 29/WG 11/N3964,N3966, "Multimedia Description Schemes (MDS) Group", March 2001, Singapore.

[11] N. Doulamis and A. Doulamis, "Optimal Content-based Video Decomposition for Interactive Video Navigation over IP-based Networks," *IEEE Trans. on CSVT* (accepted for publication) .

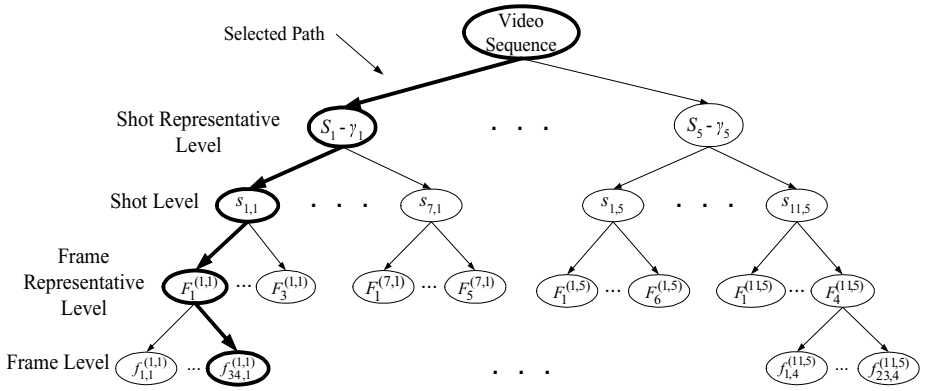


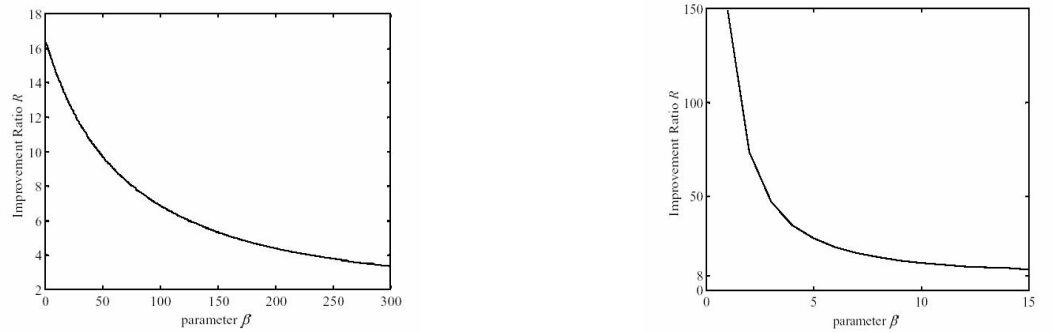
Figure 1. An example of a decomposition tree.

Table I. The Classification accuracy as well as the improvement ratio obtained for different clustering methods.

Classification Methods	Classification Accuracy	Improvement Ratio
The Method of [11]	78%	84.44
The Method of [5]	75%	78.20
The Method of [7]	72%	71.47

Table II. The Improvement ratio for different summarization methods.

Summarization Methods	Improvement Ratio
The Method of [10]	67.42
The Method of [3]	20.18
The Method of [8]	42.30
The Method of [7]	24.20
The Method of [5]	26.13



(a) (b

Figure 2. The improvement ratio R with respect to the parameter β , which regulates the distribution of the probabilities c_i , b_i , r_i .