HIERARCHICAL MOTION ANALYSIS FOR FAST SUMMARISATION OF SCALABLE CODED VIDEO

Marta Mrak[*], *Janko Ćalić*[*], *Giovanni Cordara*[+] *and Ahmet Kondoz*[*]

[*] I-Lab, Centre for Communication System Research
University of Surrey
Guildford, United Kingdom

[+] Multimedia Technologies
Telecom Italia LAB
Torino, Italy

ABSTRACT

Due to a high demand for efficient video summarisation and video adaptation technologies, this paper focuses on utilisation of compressed domain feature extraction and hierarchical analysis of motion information in scalable video in order to generate intuitive visual summaries. By combining the analysis of inherently hierarchical motion activity measure and a fast geometrical curve simplification algorithm, a set of the most representative key-frames is generated in a very fast and robust manner. The experimental results show good subjective representation, while the method efficiency enables fast generation of summaries of large-scale video repositories.

***Index Terms*—**video analysis, video summarisation, hierarchical motion coding

1. INTRODUCTION

Having experienced a proliferation of video compression algorithms that provide high compression of the digital content and thus its easy accessibility, a number of applications supporting sharing, browsing and creation of new digital content have become available. With this rapid growth of available content and related services, fast video summarisation techniques have become highly desirable. Their aim is to provide intuitive video representation in an efficient and robust manner. The existing video summarisation algorithms provide a selection of representative video frames by analysing perceptual features of video content. However, reliable video analysis systems are domain oriented and are not commonly used in commercial applications due to a high complexity and memory requirements of analysis for large-scale video databases.

In order to overcome the computational and memory demands, the compressed domain video analysis approach utilises features that can be extracted directly from the compressed domain video stream [1-3]. Furthermore, new flexible video compression concepts based on scalable video coding [4-6], offer possibilities for inherently hierarchical analysis of videos.

In this paper, we present a novel algorithm for video summarisation based on the compressed domain analysis from the scalable coded videos. In Section 2, a video analysis methodology that exploits compressed domain features is presented. A hierarchical analysis of videos which follows the layered structure of scalable bit-streams is proposed in Section 3. Key-frame selection algorithm is presented in Section 4, while the experimental results and final conclusions are given in Section 5 and Section 6, respectively.

2. VIDEO ANALYSIS

In video compression the temporal decorrelation of video frames is performed using temporal decomposition driven by motion information. The motion information captures the main temporal properties of videos and can be used for video analysis. The layered structure of scalable video enables independent access to certain portions of the bit-stream which can be exploited for fast video analysis. In the following subsections the properties of a layered bit-stream are explained and a metric for extraction of relevant features is devised.

2.1. Temporal video decompositions

Motion compensation in video coding enables exploitation of temporal redundancies in order to achieve high compression rates. It results in temporally decorrelated frames and motion information which describes the compensation process. The process in which the frames are motion compensated is often called temporal decomposition. It can be characterised by several decomposition levels which lead to hierarchical video representation. Such schemes are used in both wavelet-based coding [4], as well as in standard video coding [6] with different levels of hierarchical B-frames.

An example of temporal decomposition is shown in Figure 1. Aiming at dyadic temporal scalability the motion compensation follows a pattern that compensates every second frame at each level of temporal decomposition. The compensated frames, depicted as H frames in the Figure 1,

The work presented was developed within VISNET II, a European Network of Excellence, funded under the European Commission IST FP6 programme.

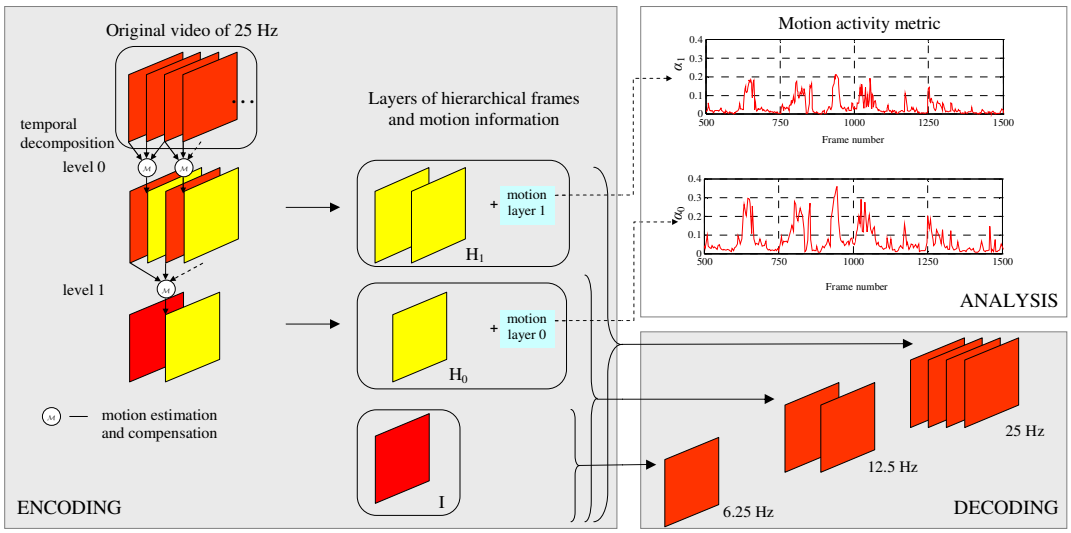

Figure 1 Motion compensation and bit-stream layers with 2 levels of temporal decompositions and 3 decoding points and the related analysis and decoding connections.

and the related motion information obtained at each level of temporal decomposition form one bit-stream layer. Motion information consists of the motion fields (motion blocks and motion vectors) associated with each frame. The uncompensated frames are subject to a higher level of temporal decomposition. Finally, after temporal decomposition the uncompensated frames, depicted as I frames in the Figure 1, form the lowest layer of scalable bit-stream.

The lower motion layers, associated with higher temporal decomposition levels, carry information obtained between more distant frames compared to the higher motion layers. Therefore it can be assumed that the most precise motion information, from the analysis point of view, is the one associated to the level 0 of the temporal decomposition. Based on the observed relations between motion layers, a mechanism for its reuse in the analysis is proposed in Section 3. The analysis algorithm uses an activity metric obtained from the frames motion fields, as defined in the following subsection.

2.2 Activity metric extraction

For selected motion compensated frames an activity metric is extracted using related motion field. Motion activity descriptor is a standard tool for capturing intensity of action correlated to human perception [3]. In the following text the extraction of the activity metric is defined for motion models with variable block sizes.

The activity metric is based on motion vectors related to each inter predicted block $\mathbf{B}_i$. For unidirectional prediction, the magnitude of the related motion vector is selected. In a case of bidirectional prediction, an average of related motion vector magnitudes is taken. The activity metric α of each frame is computed as the weighted standard deviation of all inter predicted blocks, whose number is N_Φ, with indices $i = 0,..., N_\Phi - 1$:

$$\alpha = \left(\frac{\sum_{i=0}^{N_\Phi - 1} \left(w(i) \cdot \left(m(i) - \overline{m} \right) \right)}{N_\Phi \cdot \sum_{i=0}^{N_\Phi - 1} w(i)} \right)^{1/2},$$

where $m(i)$ corresponds to averaged motion vector magnitude for i-th block and $\overline{m}$ is average magnitude for given frame. The average magnitude is computed as:

$$\overline{m} = \frac{\sum_{i=0}^{N_\Phi - 1} \left(m(i) \cdot w(i) \right)}{N_\Phi \cdot \sum_{i=0}^{N_\Phi - 1} w(i)}.$$

The weighting factor $w(i)$ of block $\mathbf{B}_i$ corresponds to its size. Weighting has been introduced to support variable size block models which are used in most efficient video codecs.

While the activity metric can be extracted for each motion compensated frame at any level of hierarchical motion compensation level, the proposed analysis algorithm takes into account temporal properties of the activity metric in order to achieve lower complexity.

3. HIERARCHICAL ACTIVITY ANALYSIS

Targeting fast selection of representative frames from a compressed bit-stream, the activity metrics associated to the video frames are assessed across motion layers. The analysis starts at the lowest motion layer $\tau = 0$ and a selective refinement of activity metrics is performed towards the higher layers with a goal to detect significant values of the metric which are not present at the current layer. Index of the highest layer is $\tau = T - 1$, where T is the total number of motion layers in given video.

The set of activity metrics that is used for the final selection of frames is obtained using hierarchical selection algorithm. Firstly, the initial selection is obtained from the lowest motion layer: $\boldsymbol{a}^0 = \boldsymbol{\alpha}^0$, where $\boldsymbol{a}$ is the vector of selected values and $\boldsymbol{\alpha}^0$ is the vector of activity metrics at $\tau = 0$. Secondly, activity metrics at higher layers (starting with $\tau = 1$) are compared with the previous selection as:

$$a^\tau(t) = \begin{cases} \alpha^\tau(t), & \text{if } \alpha^\tau(t) > \alpha^{\tau-1}\left(\lfloor t/2 \rfloor\right) \cdot k^\tau \cdot \gamma^\tau \\ a^{\tau-1}\left(\lfloor t/2 \rfloor\right) \cdot k^\tau, & \text{otherwise} \end{cases}, \quad (1)$$

where $\gamma^\tau > 1$ is a weighting factor which controls the selection process, k^τ is a normalisation factor between different scales and $\lfloor \cdot \rfloor$ represents rounding to the nearest lower integer value. The normalisation factor is defined as $k^\tau = \bar{\alpha}^\tau / \bar{\alpha}^{\tau-1}$, where $\bar{\alpha}^\tau$ correspond to expected mean value of activity metrics at given level. The dimension of selected value vector $\boldsymbol{a}^\tau$ doubles with each higher scale, matching the size of $\boldsymbol{\alpha}^\tau$.

For $\tau = 1$ all $a^1(t)$ are subject to comparison. Analysis on higher layers is performed only on samples that were already updated with higher scale values while the other values at that level are only normalised as $a^\tau(t) = a^{\tau-1}(\lfloor t/2 \rfloor) \cdot k^\tau$. This bottom-up selection of activity metrics provides hierarchy between motion layers with detection of maxima at immediate higher scales. The final vector of obtained activity metrics is denoted as $\boldsymbol{a}$ and is a subject of further analysis which determines the key-frames for summarisation, as defined in the following section.

4. SUMMARISATION BY METRIC SIMPLIFICATION

In order to locate the most representative frames in the given video sequence, the final activity metric $a(t)$ is analysed using the Discrete Curve Evolution (DCE), a geometrical shape simplification algorithm that leads to the simplification of curve complexity with no peak rounding effects and no dislocation of relevant features [7].

The curve evolution process is an iterative algorithm guided by a relevance measure, which needs to be stable with respect to noisy deformations. The DCE algorithm initialisation creates a vector $\mathbf{A}$ that comprises ordered pairs of activity metric values and their positions in the video:

$$\mathbf{A} = \left[\mathrm{A}_0, \mathrm{A}_1, ..., \mathrm{A}_{N-1}\right] = \left[\left(a_0, t_0\right), \left(a_1, t_1\right), ..., \left(a_{N-1}, t_{N-1}\right)\right]$$

where the values are set to $a_l = a(l)$ and $t_l = t(l)$ as obtained from the hierarchical analysis.

At each iteration step j, starting from $j = 0$, the element of vector $\mathbf{A}$ with minimal relevance measure is removed, thus incrementally simplifying the motion activity metric. The index r of the element to be removed is chosen as

$$r = \arg\min_i \left(k_i\right)$$

where $k_i = (a_i - a_{i-1}) \cdot (t_i - t_{i-1}) - (a_{i+1} - a_i) \cdot (t_{i+1} - t_i)$ is the chosen relevance measure that is proportional to a change of area below the motion activity curve caused by the removal of the point i on the curve. The elements of new activity metric vector $\mathbf{A}'$ are obtained as:

$$\mathrm{A}'_i = \begin{cases} \mathrm{A}_i, & i < r \\ \mathrm{A}_{i-1}, & \text{otherwise} \end{cases},$$

for each $i = 0,..., n - 2$, where $n = \dim(\mathbf{A}) = N - j$.

The final stage of DCE algorithm is achieved once the required number of local minima has been reached in simplified motion activity curve. The number of local minima corresponds to the number of frames required for the summary. Being located at the local minima of motion activity, the key-frames will have maximum probability of avoiding motion blur and other artefacts due to object motion or camera work.

5. EXPERIMENTS

The experimental tests have been performed on TRECVID 2006 content for evaluation on video retrieval systems. Scalable video coding has been performed with 6 levels of temporal decomposition on 25 Hz videos which corresponds to I frame period of 2.56 s. In addition to high compression, this decomposition pattern produces a high number of motion layers that can be independently accessed for the analysis purposes. The length of test sequences is in the range of 3100 to 3500 frames and the target video summary length is 8 frames which corresponds to approximately 2 frames per one minute of video in average.

Two different analysis modes have been tested, both based on the proposed motion activity measure and DCE. The first (reference) mode of the analysis does not use hierarchical selection of activity values. In this mode the activity metric is computed from selected motion layer on which the frames are selected using DCE. The second mode is the proposed hierarchical mode in which the refinement of motion activity vector is performed from lower to higher scale.

The results of summarisation for 2 test sequences (S1 - "FRANC101" and S2 - "MRS042546") are presented in Figure 2. For both test sequences, the analysis algorithms have achieved representative summaries. Evaluation of the quality of summarisation and related complexities are shown in Table 1. The quality of summarisation has been evaluated using user-provided intervals of test videos from which a key frame should be used in the summarisation. Such user expectations are used as a ground truth. The number of intervals depends on the subjective perception of content and not on the number of frames requested from the summarisation algorithm. Both summarisation methods achieved the same quality of about 80 % of expected intervals captured, as presented in the Table 1 in the user expectation column.

a) summaries based on single layer analysis

b) summaries based on hierarchical analysis

Figure 2 Results of the video summarisation using motion activity measure from a) single motion layer and b) hierarchical analysis.

Table 1 Numerical evaluation of summarisation results.

sequence	User expectation		Complexity	
	Single layer	Hierarchical analysis	Single layer	Hierarchical analysis
S1	83 %	83 %	12.5 %	2.4 %
S2	80 %	80 %	12.5 %	3.2 %

Compared to the algorithm that uses analysis of video from information at one motion layer, the proposed hierarchical algorithm provides significant complexity reduction. The complexity is evaluated in terms of a ratio of frames used in activity metric extraction and overall number of frames. From results presented in the Table 1, it can be observed that the proposed hierarchical analysis requires on average 4.5 times lower computational complexity. Lower complexity is achieved by hierarchical analysis that overcomes the computation of motion activity for all frames.

From Figure 2 it can be seen that despite high matching to user expectation, both summarisation algorithms in some cases selected more that one frame from a single user-defined video interval. This mismatch naturally occurs when the sequence has to be represented with predefined number of frames. Its negative influence to the overall user experience can be minimised by application of advanced layout optimisation algorithm that uses frame clustering and variable size frame representation, as described in [8].

6. CONCLUSIONS

This paper presents a novel method for video summarisation by utilising a hierarchical analysis of compressed domain features using scalable video coding technology. By direct extraction of motion information from video stream compressed using scalable video coding, motion activity features are extracted in a very efficient manner. Furthermore, the inherent multi-scale character of the motion activity metric is exploited to apply a hierarchical analysis for video summarisation. Finally, using a curve simplification methodology, a set of representative key-frames is extracted in order to synthesise a video summary. Results demonstrated that the highly efficient feature extraction and analysis algorithm can be used for generation of compact yet informative summaries of scalable video. The future work will focus on exploiting the low complexity of this algorithm in order to generate real-time interactive interfaces for large-scale video databases.

REFERENCES

[1] V. Kobla, D. S. Doermann, K.-I. Lin, and C. Faloutsos, "Compressed-domain video indexing techniques using DCT and motion vector information in MPEG video", *in Proc. SPIE 97*, Vol. 3022, pp. 200 - 211, 1997.

[2] J. Bescos, J. M. Martinez, L. Herranz, and F. Tiburzi, "Content-driven adaptation of on-line video", *Signal Processing: Image Communications*, No. 22, pp. 651 - 668, 2007.

[3] A Rosenfeld, D. Doermann, D. DeMenthon (Editors), *Video Mining,* Kluwer Academic Publishers, 2003.

[4] N. Adami, A. Signoroni, and R. Leonardi, "State-of-the-art and trends in scalable video compression with wavelet-based approaches", *IEEE Trans. on Circ. and Sys. for Video Tech.*, Vol. 17, Iss. 9, pp. 1238 - 1255, Sept. 2007.

[5] M. Mrak, N. Sprljan, and E. Izquierdo, "Motion estimation in temporal subbands for quality scalable motion coding", *Electronics Letters*, No. 41, pp. 1050 - 1051, 2005.

[6] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. on Circ. and Sys. for Video Tech.*, Vol. 17, Iss. 9, pp. 1103 - 1120, Sept. 2007.

[7] L. J. Latecki and R. Lakamper, "Convexity rule for shape decomposition based on discrete contour evolution", *Computer Vision and Image Understanding*, Vol. 73, pp. 441–454, 1999.

[8] J. Calic, D. P. Gibson, and N. W. Campbell, "Efficient layout of comic-like video summaries", *IEEE Trans. on Circ. and Sys. for Video Tech.*, Vol. 17, Iss. 7, pp. 931 - 936, July 2007.