A MULTIRESOLUTION TECHNIQUE FOR VIDEO INDEXING AND RETRIEVAL

Janko Calic and Ebroul Izquierdo
Department of Electronic Engineering, Queen Mary, University of London
Mile End Road, E1 4NS London, UK,
{janko.calic, ebroul.izquierdo}@elec.qmul.ac.uk

ABSTRACT

This paper presents a novel approach to the multiresolution analysis and scalability in video indexing and retrieval. A scalable algorithm for video parsing and key-frame extraction is introduced. The technique is based on real-time analysis of MPEG motion variables and scalable metrics simplification by discrete contour evolution. Furthermore, a hierarchical key-frame retrieval method using scalable colour histogram analysis is presented. It offers customisable levels of detail in the descriptor space, where the relevance order is determined by degradation of the image, and not by degradation of the image histogram. To assess the performance of the approach several experiments have been conducted. Selected results are reported in this paper.

1. INTRODUCTION

Having an indisputable success, development of visual content-based indexing and retrieval (CBIR) systems has settled down offering a wide spectrum of low-level perceptual content representation. However, the scalability problem remains open. Lack of video analysis at the desired level of detail and absence of scalable hierarchical indexing and retrieval methods were the main inspiration for this work.

First task of a CBIR system is to parse video into temporal semantically coherent units and to extract a set of frames that represent the visual semantics of the sequence in the best way. Considering the fact that video is nowadays mainly stored in compressed form, we will focus on video parsing algorithms that utilise features extracted directly from compressed video streams (MPEG, H.26X). Yeo and Liu in [1] presented an algorithm that analyses a sequence of reduced images extracted from low frequency DCT coefficients called the DC sequence. Lee et al. [2] exploited information from the few additional AC coefficients in the transformation domain to extract and analyse binary edge maps. Several authors proposed utilisation of the MPEG motion compensation variables, having the inherent measure of frame similarity and sequence continuity and being directly embedded in MPEG stream. Pei et al. [3] analysed patterns of macroblock types to detect shot changes, while Kobla et al. [4] introduced a complete indexing engine that analyses only compressed domain motion features.

Extracting a key-frame to represent the semantics of the whole shot appeared to be the first, rather modest, approach to semantic analysis in conventional CBIR systems. Criteria to do that was either by tracking the temporal variation of the visual features where the number of key-frames was determined a posteriori [5] or by predefining the number of key-frames [6]. Apparently, none of the published works have solved essential requirements of real-time processing capability and scalability.

The main objective of this work is to develop a scalable and highly efficient technique for indexing and retrieval of the key-frame images by using only colour features. A family of successively simplified image features is defined using colour histograms. The image colour distribution is used to define an image dependant quantisation according to the visual information loss at a given scale. This novel approach offers customisable number of levels of detail, where the relevance order is driven by uniform degradation of the image, and not by uniform degradation of the image histogram. Conventional strategies either extract descriptors from multi-resolution image space [7], which is demanding in terms of resources, or quantising the descriptor space independently of the image itself [8]. The generated family of histograms is then used to define descriptors at various levels of detail. Using the hierarchical descriptor structure irrelevant details and noise are removed in a very early processing step. Large sets of non-similar images are discriminated at very low computational cost using low detailed descriptions. The search is then refined progressively until only a few of very similar objects or images are found and ranked using higher levels of detail.

This paper is organized as follows. In Section 2 the scalable method for event detection and key-frame extraction is presented. Section 3 describes the colour indexing algorithm, as well as introduces novel approach to hierarchical histogram quantisation. Overall results are presented in Section 4, while the Section 5 brings final conclusions and a summary of the paper.

2. KEY-FRAME EXTRACTION METHOD

Since the MPEG sequence has a high temporal redundancy within a shot, a continuously strong inter-frame reference will be present in the stream as long as no significant changes occur in the scene [9]. The "amount" of inter-frame reference in each frame and its temporal changes can be used to define a metric, which measures the probability of scene change in a given frame. We propose to extract only macroblock (MB) type information from the MPEG stream and, by analyzing it, to measure this "amount" of inter-frame reference.

Without loss of generality we assume that a *Group Of Pictures* (GOP) in the MPEG stream will have the standard structure [IBBPBBPBBPBBPBB], enabling us to analyse the triplet frame structure: $R_1\ B_2\ b_3\ \ R_4\ B_5\ b_6\ \ldots\ R_i\ B_{i+1}\ b_{i+2}\ \ldots$ In the sequel, both types of the reference frames (I or P) are denoted as R_i, front bi-directional frame of the triplet as B_i, while the second bi-directional frame is denoted as b_i.

If the first referenced frame B_i is the first frame with different visual content, the next reference frame R_{i+2} predicts backwards a significant percentage of MBs in both B_i and b_{i+1}. If the content change occurs at the reference frame R_i, then the bi-directional frames B_{i-2} and b_{i-1} will be mainly predicted forwards by the previous reference frame R_{i-3}. Finally, if the content change occurs at b_i, then B_{i-1} will be strongly predicted forward by the previous reference frame $R_{i-2,}$ while b_i will be predicted backwards by the next reference frame R_{i+1}.

Let $\Phi_T(i)$ be the set containing all forward referenced MBs and $B_T(i)$ the set containing all backward referenced MBs in a given frame with index i and type T. In the same manner, we define sets of intra coded MBs as $I_T(i)$ and interpolated MBs as $\Pi_T(i)$. Then we denote the cardinalities of the corresponding sets as: $\varphi_T(i)$, $\beta_T(i)$, $\iota_T(i)$ and $\pi_T(i)$. The metric $\Delta(i)$ used to determine a visual difference measure within a frame triplet is defined as:

$$\Delta(i) = k_{\varphi B}\varphi_B + k_{\varphi b}\varphi_b + k_{\beta B}\beta_B + k_{\beta b}\beta_{bB} + k_{\iota B}\iota + k_{\iota b}\iota_b + k_{\pi B}\pi_B + k_{\pi b}\pi_b$$

By analysing the prediction character and behaviour in one frame triplet, we can estimate the changes in visual content within. Depending on the frame type T(i), there are three different linear combinations of variables $\varphi_T(i)$, $\beta_T(i)$, $\iota_T(i)$ and $\pi_T(i)$ for both bi-directional frames in a frame triplet. Each linear combination has two main coefficients that are directly proportional to the visual content change within predicted and reference frame in a frame triplet (k=+1), and two that are inversely proportional (k=-1) to it. Additional factors k_* and $k_.$ are describing overall change in a triplet, one in direct ($k_.$=0.5) and one in inverse (k_*=-0.5) proportion.

Since the metrics value is determined separately for each frame and the content change is based on frame triplet element low-pass Gaussian filtering with kernel proportional to triplet length is applied to eliminate the noise.

In order to extract a number of representative frames from the sequence the previously defined difference metrics $\Delta(i)$ is simplified in a way that spurious and small changes in the metrics curve are discarded without any influence on the main features of the difference metrics. The algorithm that has these features is Discrete Curve Evolution (DCE) [10]: (i) It leads to the simplification of curve complexity with (ii) No peak rounding effects and no dislocation of relevant features and (iii) The relevance measure K is stable with respect to noisy deformations. Flowchart of DCE algorithm is depicted in Figure 1.

Let $D_m = s_0, \ldots, s_{m-1}$ be a decomposition of a digital curve S into consecutive digital line segments. The algorithm computes the decomposition D_k for each stage of the discrete curve evolution k>3 until we reach wished number of key points (NOKP). Number of key-frames (NOKF) after the DCE algorithm can be determined either a priori or *a posteriori*, by defining detection sensitivity. The input video sequence has NOF frames.

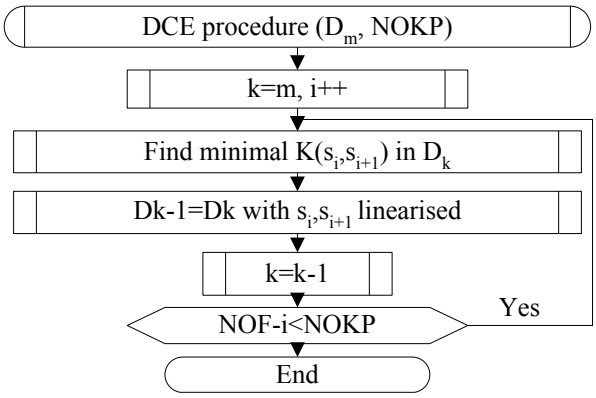

Figure 1. Flowchart of the DCE algorithm

For each two adjacent line segments s_1, s_2 in the decomposition of a digital curve S, we determine the relevance measure $K(s_1, s_2)$, which represents the significance of the contribution of arc $s_1 \cup s_2$ to the shape of S. The value $K(s_1, s_2)$ can be interpreted as the cost required for linearization of arc $s_1 \cup s_2$. In this case, the linearization cost $K(s_1,s_2)$ of any supported arc s_1,s_2 depends on its length, its global curvature and area below the arc. Let s_1=AB and s_2=BC be two consecutive line segments in the decomposition of curve S, so that $\beta=\alpha_1+\alpha_2$ is the turn angle. The corresponding cost function $K(s_1, s_2)$ is given by the equation:

$$K(s_1, s_2) = \left| \beta(s_1, s_2) \cdot (l_1 + l_2) \cdot P_{\Delta(ABC)} \right|$$

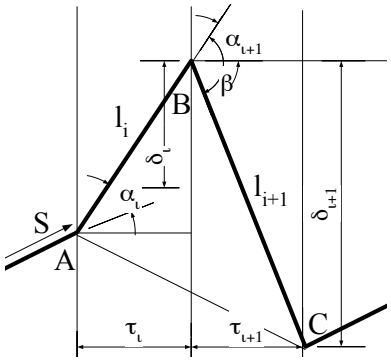

Figure 2. DCE linearization

Observing an arc linearization example given in Figure 2, formulae for each element in equation above for relevance measure are given as:

$$\delta_i = \Delta(i+1) - \Delta(i) \quad , \quad l_i = \sqrt{\tau_i^2 + \delta_i^2}$$

$$\beta(s_i, s_{i+1}) = acrtg(\delta_i / \tau_i) - acrtg(\delta_{i+1} / \tau_{i+1})$$

$$P_{\Delta ABC} = 1/2 \cdot (\delta_i \tau_i + \delta_{i+1} \tau_{i+1})$$

Key-frames positions are determined by locating the local minima in simplified metrics curve, while shot change central points are located as the local maxima.

3. COLOUR INDEXING AND HIERARCHICAL HISTOGRAM QUANTISATION

Among different colour representations the HLS model has two important characteristics: it is easy to use and it produces colour components that closely follow those perceived by humans [11]. For this reason a family of quantised colour histograms in the HLS colour space is used as the set of the image descriptors.

To generate a hierarchical family of histograms, a continuous histogram simplification algorithm is implemented. In each step of the algorithm the least significant colour component is removed and the image degradation measure is calculated. By reaching the desired measure of image degradation the representing histogram is extracted as the image descriptor at that particular level of detail.

The simplification algorithm is very similar to the DCE algorithm applied to the frame difference metrics, with modified relevance measure and without segment linearization, but simple colour component removal. The relevance measure function in this case is defined as:

$$K(i) = hist(i) \cdot \log(\tau_i + \tau_{i+1})$$

where *hist(i)* is the image histogram and τ_i is interval between components *i* and *i-1*. The algorithm removes the colour component with the lowest relevance measure value.

The image degradation function Df(n) at the algorithm step n is defined as the cumulative sum of the previously removed histogram bin values:

$$Df(n) = \sum_{\forall m, hist(m) \in hist'} hist(m)$$

where *hist'* is a set of previously removed components. The value of *Df(n)* is equal to the number of image pixels that got removed during the histogram simplification process. The user can predefine the levels of the image degradation according to the addressed application.

To measure the colour similarity between key-frames at a given scale, the Hausdorff metric is applied. Each histogram is represented by a set of points $A = \{p_1, p_2, \ldots, p_k\}$ for $k \in [0, 360)$. The distance from any $p \in A$ to another set $B = \{q_1, q_2, \ldots, q_l\}$ is defined as:

$$d(p, B) = \min_{q \in B} \|p - q\|$$

The directed Hausdorff distance from *A* to *B* is given by:

$$hdist(A, B) = \sum_{p \in A} d(p, B).$$

Using that, the final distance between *A* and *B* is:

$$D(A, B) = hdist(A, B) + hdist(B, A).$$

4. RESULTS

Some of the used test sequences were produced by the Multimedia & Vision Research Lab, Queen Mary, University of London, while others were provided by Computer Vision Department, Dublin City University, Dublin, Ireland.

The statistical performance evaluation of temporal segmentation of the video sequences is "based on the number of missed detections (MD's) and false alarms (FA's), expressed as recall and precision" [12]:

$$Recall = \frac{Detects}{Detects + MD's}, \quad Precision = \frac{Detects}{Detects + FA's}$$

Manually detected positions of the shot boundaries were used as the ground truth. The number of missed detections and false alarms were based on that information. There were three main categories of video material analysed: news, soaps and commercials.

The shot changes detection procedure showed excellent results for different types of changes, as shown in Table 1.

	Detect	Missed	False	Recall	Prec.
News	87	2	6	98%	94%
Soap	92	2	9	98%	91%
Comm	127	9	16	94%	88%

Table 1. Shot changes detection results

After the number of experiments with different abstraction rates and different video content, the conclusion is that the key-frame extraction algorithm shows good visual content summarising results that can be used to perform high-level semantic analysis. Figure 3 shows the curve evolution of the metric at the different level of detail. This evolution corresponds to a short commercial clip.

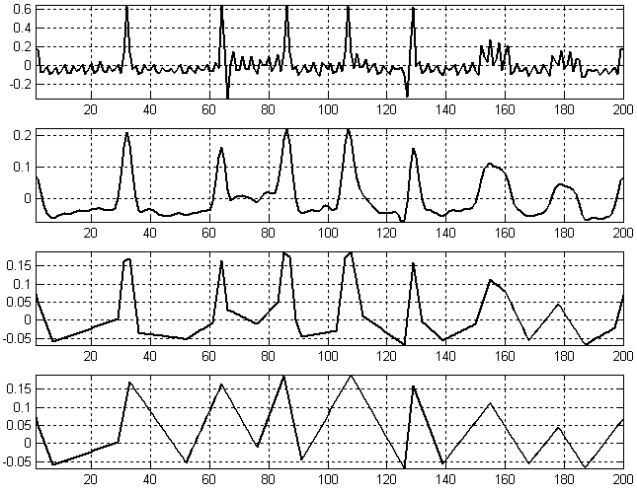

Figure 3. DCE algorithm results

It can be seen that the DCE algorithm removes less significant changes step by step without dislocating the vertices of the original curve.

A database with population of about 1200 images has been used for retrieval evaluation. The implemented graphical user interface comprises several functionalities: key-frame browsing by subjective similarity, relevance feedback, spatial region analysis, etc. Query by example evaluation showed excellent results both in terms of visual and content similarity.

5. CONCLUSIONS

A novel approach to the scalability issues in CBIR systems is proposed. First, a scalable algorithm for temporal video parsing and key-frame extraction that uses statistics of the MPEG motion variables was introduced. Experimental results presented in Section 4 show high accuracy and robustness with real-time processing capability and completely scalable analysis. We are investigating possibilities of improving the quality of key-frame presentation and introducing the high-level semantics descriptors.

A scalable and efficient technique for key-frame image indexing and retrieval using colour histograms has been also developed. A family of quantised histograms that generate the descriptor space is arranged by a criterion based on image degradation. The implemented system allows content-based search and adaptively weighted relevance feedback. Currently the comparisons between the results obtained with the implemented system and other quantization techniques are performed to measure the effectiveness of the proposed approach. Further research should consider self-learning ability and a more advanced interpretation of the relevance feedback. Additional visual primitives derived from texture and shape should also be included in the final system. Additional criteria for both image degradation measure and histogram relevance measure should be further investigated (e.g. Earth Movers Distance).

ACKNOWLEDGEMENTS

The research leading to this paper has been supported from the UK EPSRC, project Hierarchical Video Indexing Project, grant number R01699/01.

6. REFERENCES

[1] Boon-Lock Yeo, Bede Liu, “Rapid scene analysis on compressed video”, IEEE Transactions on Circuits & Systems for Video Technology, vol.5, no.6, Dec. 1995, pp.533-44. Publisher: IEEE, USA

[2] Seong-Whan Lee, Young-Min Kim, Sung Woo Choi, “Fast scene change detection using direct feature extraction from MPEG compressed videos”, IEEE Transactions on Multimedia, vol.2, no.4, Dec. 2000, pp.240-54. Publisher: IEEE, USA

[3] Soo-Chang Pei, Yu-Zuong Chou, ”Efficient MPEG compressed video analysis using macroblock type information”, IEEE Transactions on Multimedia, vol.1, no.4, Dec. 1999, pp.321-33. Publisher: IEEE, USA

[4] V. Kobla, and D. Doermann, "Indexing and Retrieval of MPEG Compressed Video", Journal of Electronic Imaging, Vol. 7(2), pp. 294-307, April, 1998.

[5] H. Zhang, A. Kankanhalli and W. Smoliar, “Automatic partitioning of full-motion video”, Multimedia Systems, vol.1, no.1, pp. 10-28, 1993.

[6] A. Hanjalic and R.L. Langendijk, "A New Key-Frame Allocation Mathod for Representing Stored Video Streams", Proc. of 1st Int. Workshop on Image Databases and Multimedia Search, 1996.

[7] Ravela S, Manmatha R, Riseman EM. "Image retrieval using scale-space matching", Proc. of ECCV '96., vol.1, 1996, pp.273-282

[8] Wu P, Manjunath BS, Shin HD. "Dimensionality reduction for image retrieval" Proc. ICIP 2000, vol.3, 2000, pp.726-9

[9] J. Calic and E. Izquierdo, "Temporal Segmentation of MPEG video streams", to be published in Image Analysis for Multimedia Interactive Services, Journal on Applied Signal Processing, 2001

[10] Latecki LJ, Lakimper R., "Convexity rule for shape decomposition based on discrete contour evolution", Computer Vision & Image Understanding, vol.73, no.3, March 1999, pp.441-54, Academic Press, USA.

[11] Swain, M. J. and Ballard, D. H. “Colour indexing,” Intl. J. of Computer Vision, Vol. 7, No. 1, pp. 11-32, 1991.

[12] Gargi U., Strayer S., “Performance Characterisation of Video-Shot-Change Detection Methods”, IEEE Trans. on Circuits and Systems for Video Technology, Vol.10, No.1, February 2000.