Signal Processing: *Image Communication* 25 (2010) 450–465

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Scalable object-based video retrieval in HD video databases

Cl. Morand [a], J. Benois-Pineau [a,*], J.-Ph. Domenger [a], J. Zepeda [b], E. Kijak [c], Ch. Guillemot [b]

[a] *LABRI UMR 5800 - Universités Bordeaux - CNRS, 351 cours de la Libération F-33405 Talence Cedex, France*
[b] *INRIA, Centre Rennes - Bretagne Atlantique, Campus de Beaulieu. F-35042 Rennes Cedex, France*
[c] *Université de Rennes 1, IRISA, Campus de Beaulieu. F-35042 Rennes Cedex, France*

ARTICLE INFO

Article history:
Received 18 September 2009
Accepted 24 April 2010

Keywords:
HD video
Scalable video object extraction
Object-based indexing
Video retrieval

ABSTRACT

With exponentially growing quantity of video content in various formats, including the popularisation of HD (High Definition) video and cinematographic content, the problem of efficient indexing and retrieval in video databases becomes crucial. Despite efficient methods have been designed for the frame-based queries on video with local features, object-based indexing and retrieval attract attention of research community by the seducing possibility to formulate meaningful queries on semantic objects. In the case of HD video, the principle of scalability addressed by actual compression standards is of great importance. It allows for indexing and retrieval on the lower resolution available in the compressed bit-stream. The wavelet decomposition used in the JPEG2000 standard provides this property. In this paper, we propose a scalable indexing of video content by objects. First, a method for scalable moving object extraction is designed. Using the wavelet data, it relies on the combination of robust global motion estimation with morphological colour segmentation at a low spatial resolution. It is then refined using the scalable order of data. Second, a descriptor is built only on the objects extracted. This descriptor is based on multi-scale histograms of wavelet coefficients of objects. Comparison with SIFT features extracted on segmented object masks gives promising results.

1. Introduction

With exponentially growing quantity of video content in various formats, including the popularisation of HD (High Definition) video and cinematographic content, the problem of efficient indexing and retrieval in video databases becomes crucial. After the success of early indexing and retrieval attempts on video considering global key-frame descriptors [1,2], the main stream of research nowadays focus on indexing and retrieval of image and video content with sparse local features. One of the first noticeable papers in this area introduced the ScavFT (scale variant feature transform) [3] for tracking purposes. The characteristic points were detected in a single video frame based on the search of a strong minimal eigenvalue of a local gradient matrix. Then this technique was extended to the temporal dimension for video retrieval in [4]. Nevertheless, for indexing and retrieval purposes, the mostly used sparse feature points and associated features, invariant to luminance, scale and rotation changes, were the SIFT proposed in a fundamental work by Lowe [5]. Since then various improvements of this approach came out and a good survey of these techniques is presented in [6].

Features sparsely distributed in the frames do not convey semantics in terms of generic objects evolving in video. Hence, several recent research works try to link the sparse features with spatio-temporal objects. Thus in [7] the MSER (Maximally Stable Extremal Regions, [8]) together with SIFT features computed on them are used to train object models in video and recognize them.

* Corresponding author.
E-mail addresses: morand@labri.fr (Cl. Morand), jenny.benois@labri.fr (J. Benois-Pineau), domenger@labri.fr (J.-P. Domenger), jzepeda@irisa.fr (J. Zepeda), ekijak@irisa.fr (E. Kijak), cguillemot@irisa.fr (Ch. Guillemot).

0923-5965/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2010.04.004

Despite the originality of these ideas and the relatively good performance of the proposed methods, these approaches do not start with an explicit object extraction from video sequences. In some sense they try to avoid solving the classical "chicken and egg" problem of semantic object extraction from video.

At the same time, the significant effort of research community related to the elaboration of MPEG4 [9], MPEG7 [10] and JPEG2000 standards [11] was devoted to the development of automatic segmentation methods of video content to extract objects. The results of these methods, e.g. [12–14], while not always ensuring an ideal correspondence of extracted object borders to visually observed contours, were sufficiently good for visual recognition of extracted object areas for fine-tuning of encoding parameters and for content description.

Hence, we are strongly convinced that the paradigm consisting in segmenting objects first and then representing them in adequate feature spaces for object-based indexing and retrieval of video remains the promising road to the success and a good alternative for local modelling of content by feature points.

Nowadays, it is impossible to imagine visual content exchanged and stored in a raw format. The standard resolution of HD video (1080p) or filmic content (4 Kp) make a raw video content a tremendous mass of data which cannot be handled without compression. The specifications of DCI (Digital Cinema Initiative, LLC [15]) made MJPEG2000 [16] the digital cinema compression standard. Furthermore MJPEG2000 is becoming the common standard for archiving [17] of cultural cinematographic heritage with the greater quality/compression compromise than previously used solutions. Nowadays, data coded with this standard constitute databases of audio-visual content so large that access to digital content requires the development of automatic methods for indexing and retrieval of this compressed content. The metadata resulting from indexing process are very heterogeneous between different systems and are shaped with MPEG7 [11] and Dublin Core [18]. This is why the JPSearch project [19] aims on establishing the standardize interfaces for an abstract image retrieval framework. One of the focuses of JPSearch is the embedding of metadata in image data encoded in the JPEG2000 standard. Hence the latest research works link content encoding and indexing in the same framework be it images or video [20]. In our work we also address this issue and propose object-based indexing of video content jointly with MJPEG2000 compression. The present work is in the continuation of the RI (Rough indexing) paradigm we developed for MPEG compressed video [21].

New compression standards such as MJPEG2000 or H.264/SVC [22] have the interesting property of scalability. Hence, a codestream formatted with one of the standards mentioned can be sent to different users with different processing capabilities and network bandwidths by selectively transmitting and decoding the related part of the codestream. The extracted sub-streams correspond to a reduced spatial resolution (spatial scalability), a reduced temporal resolution (temporal scalability), a reduced quality for a given spatio-temporal resolution and/or for a reduced flow (SNR scalability). This scalability property gives an exciting perspective of video retrieval at various resolution levels, hence reducing computational work load and allowing for "cross-resolution" retrieval. The contribution of our work is two-fold. In the framework of object-based "scalable" indexing and retrieval of video content we first propose an object extraction/segmentation method operating on a scalable MJPEG2000 compressed stream. Therefore, the extraction process has to be performed with the only part of the stream available after transmission, i.e. the segmentation should be made in the only one direction from low resolution to high resolution. Secondly, in the paradigm of "object-based" video content retrieval we propose object descriptors for a "scalable retrieval" of content enabling search on reduced resolution level.

Thus, in our vision, the first step in object-based indexing is the foreground object extraction. Several approaches have been proposed in the past and most of them can be roughly classified either as intra-frame segmentation based or as motion segmentation based methods. In the former approach, each frame of the video sequence is independently segmented into regions of homogeneous intensity or texture, using traditional image segmentation techniques [23], while in the latter approach, a dense motion field is used for segmentation and pixels with homogeneous motion field are grouped together [24]. Since both approaches have their drawbacks, most object extraction tools combine spatial and temporal segmentation techniques [25,26]. Challenge resides in applying this scheme without decompressing the video, as the low-level content descriptors, such as coefficients in the transform domain, can be efficiently re-used for the content analysis task [21].

In the case of the MJPEG2000 standard, one difficulty is that the standard does not provide motion descriptors and so they have to be estimated. The ME (Motion Estimation) problem in the wavelet domain has largely been studied in the literature [27]. We also proposed a first approach for motion estimation on JPEG wavelet pyramid in [28]. In this paper, we will present its improvements allowing reducing the negative effect of shift-variance property.

A second step in object-oriented indexing consists in defining features on the object effectively found. Following the RI paradigm, these features have to be defined in the compressed domain, i.e. the wavelet domain. Several indexing techniques in the wavelet domain exist for JPEG2000 compressed still images. Among them, we can cite the histogram-based techniques. A histogram is computed for each subband and comparison is made subband by subband [29]. The main drawback of such a technique is that it works only with limited camera operations. To reduce complexity and improve the robustness to illumination changes, [30] proposes modelling the histograms using a generalized Gaussian density function. Other indexing techniques are texture oriented [31]. In this work, we propose a global histogram-based descriptor for object in the wavelet domain at different levels of spatial scalability. We also compare the efficiency of this descriptor with an approach consisting in matching