Indexing Camera Motion Integrating Knowledge of the Quality of the Encoded Video

P. Krämer, J. Benois-Pineau, member IEEE, M. Gràcia Pla

***Abstract*—Fast indexing of video contents in the compressed domain has become an important task as growing quantities of multimedia (MM) digital content are available in this form. In this paper we present a method for fast indexing of camera motion of MPEG1 and 2 compressed video. We use P-frame motion vectors and extract some knowledge on the quality of the compensated motion from the compressed stream. It is then used for decision making on the motion refinement. Then camera motion is indexed in terms of physical motions. Results obtained on the TREC Video test data set are interesting.**

***Index Terms*— video indexing, camera motion, compressed streams.**

I. Introduction

Indexing and annotating large quantities of films and video material has become an increasing problem for the media industry. Today, indexing for large application areas such as broadcast, archives, and home MM devices definitely follows MPEG7 – the compliant way. This is a standard [1] for describing the multimedia content. For visual media, it defines descriptors to characterize the content on a visual basis. In video, which intrinsic property is motion, it proposes motion descriptors. Nevertheless, MPEG7 does not give hints on how to produce a standard compliant description of e.g. camera motion, and how to translate this description into features easily interpreted by humans such as tilt, zoom, or pan… A lot of multimedia content is already available in compressed form. Furthermore, a digitization of the existing video content and digital production of new content are today unthinkable without compression. Thus a lot of work [2 – 4] has been devoted to the estimation of the camera model from motion vectors contained in the compressed stream. This work is another step forward in the general framework which we call "Rough Indexing Paradigm" and has been developed since [5]. A whole lot of indexing tasks such as shot boundary detection, scene grouping, video summarization, video object extraction, or motion characterization can be fulfilled on degraded and low-resolution/low-level data produced by encoding video streams with current encoders (MPEG1, 2, H.264 …). We claim that a compressed stream is a rich source of input data for indexing and this is only the matter of interpretation for the intelligent use of it. In this paper we show how we can truly use not only MPEG (1 or 2) motion vectors, but also the information on the quality of their estimation in order to estimate the camera model (Section 2) and to qualify motion in the humanly interpretable way (Section 3). This is for instance a task of camera motion characterization in TREC Video 2005, where we did participate. We show how this knowledge helps us to improve the indexing results and give the perspectives of this work (Section 4).

P. Krämer and J. Benois-Pineau are with LABRI UMR CNRS/University of Bordeaux 1/Enseirb/INRIA laboratory, 351, crs de la Libération, 33405 Talence Cedex, France; petra.kraemer, jenny.benois@labri.fr; phone 33 5 40 00 84 24, fax 33 5 40 00 66 69. M. Gràcia Pla has been on master position in LABRI on leave from UPC, Barcelona, Spain.

II. Global motion estimation and correction from MPEG compressed video

In this section we address the problem of estimating the global (camera model) in a video sequence. Here we use motion compensation vectors from P-frames. In order to remain the same temporal resolution and get a smooth motion trajectory, we interpolate it for I-frames. Finally, as MPEG motion vectors are not computed for analysis purposes, but for optimal encoding, they can be very much erroneous (e.g. in case of strong motion), we propose how to detect such encoder failures and how to correct the motion.

II.1 Global motion estimation from P-frames

Here we rely on our previous work [5] and use a 6 parameter affine camera model. We suppose [5] that an MPEG macro-block displacement vector is expressed as:

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} x - x_g \\ y - y_g \end{pmatrix} \quad (1)$$

where $a_1, ..., a_6$ are the global motion parameters of camera and $(x_g, y_g)^T$ denotes the image center. The estimation by a robust estimator that we proposed in [5], allows classifying macro-blocks (MBs) as conformant to the model, what we call the "dominant estimation support", or outliers. The latter contain intra-coded MBs, MBs in moving objects and in occluding areas. This approach supposes that in a current P-frame, there are motion vectors, which express the apparent camera motion. Unfortunately this is not always the case. In order to re-cover the real camera motion in such frames it is necessary to detect encoder failures and to correct the motion.

II.2 Detection of frames with low–quality motion and motion correction

If the MPEG encoder motion estimator failed, the motion compensation error encoded in the MPEG stream is strong. Such failures are very much dependent on the parameter settings of the encoder and are specifically observed in the case of strong motion (e.g. soccer content).

We compute the mean low frequency energy E_t on the dominant estimation support D_t i.e. excluding the motion outliers:

$$E_t = \frac{1}{\|D_t\|} \sum_{p \in D_t} DC_P^{err}(p,t)^2 \quad (2)$$

Here $DC_P^{err}(p,t)$ are the DC coefficients extracted from the encoded error in P-frames.

To take the decision if the motion model has to be corrected, we use the temporal mean γ_t of (2). If the instantaneous value of (2) exceeds $\alpha\gamma_t$, with $\alpha \geq 1$ then the motion will be corrected.

To fulfill this correction we first interpolate the motion model from neighboring P-frames by a linear regression. This interpolation is used as the initialization of the model estimate in the gradient descent scheme.

Here we minimize the functional of the mean square error of the motion compensation at DC resolution on the dominant estimation support:

$$MSE_t = \frac{1}{\|D_t\|} \sum_{p \in D_t} \left(I_t(p) - I_{t-1}(p + \vec{d})\right)^2 \quad (3)$$

The optimization is done in the parameter space by gradient descent:

$$\Theta_t^{i+1} = \Theta_t^i - \frac{\varepsilon}{2\|D_t\|} G^i$$

with G as the gradient of (3) and ε as the adaptive gain matrix.

III. CAMERA MOTION INDEXING

The objective here is to translate the motion model (1) into physical motion, interpretable by humans, such as pan, tilt, or zoom. To do this we follow [6] and reformulate the model (1) as:

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} pan \\ tilt \end{pmatrix} + \begin{pmatrix} zoom \cdot x - rot \cdot y + hyp1 \cdot x + hyp2 \cdot y \\ zoom \cdot y + rot \cdot x - hyp1 \cdot y + hyp2 \cdot y \end{pmatrix} \quad (4)$$

Then two statistical hypotheses are tested on each parameter of this model. The first one H_0 consists in supposing that the parameter is significant, the second one H_1 assumes that the component is not significant, i.e. equals zero.

The likelihood function *f* for each hypothesis is defined with respect to the residuals between the estimated model and the MPEG motion vectors. These residuals are supposed to follow the bi-variate Gaussian law. The decision on the significance is made by a comparison of the log-likelihood ratio with a threshold. We used this scheme in our previous work, but in case of the knowledge on a bad estimation that is available from (2), we do not compute residuals between the erroneous MPEG motion vectors and those obtained by the re-estimated model. The interpolated parameters are used as reference (light correction) in this case.

IV. RESULTS AND CONCLUSION

To assess the improvement due to the proposed integration of the knowledge on erroneous motion and re-estimation of motion (3), we conducted experiments on the evaluation set of the TREC Video camera motion task http://www-nlpir.nist.gov/projects/trecvid/ in which we participated in 2005. A subset of 4 videos containing visually observable motion was chosen. Using $\alpha = 4.0$ in the decision rule, about 4% of the P-frame motion is corrected. Due to this correction we obtain a mean precision of 76% and a mean recall of 86.1%. Without the correction 74.5% and 78.7% are obtained respectively. We have to stress that the increase of recall of 8 % is already very much significant for this task.

Hence in this paper we proposed a new method for motion correction when estimating and indexing camera motion from compressed (MPEG1 and MPEG2) video streams.

We tested it for indexing purposes on the MPEG1 compressed TREC Video test set. For video summarizing by mosaicing from compressed streams and for other indexing applications (shot boundary detection, object extraction) we work on MPEG2 compressed streams as well. There is no principal difference and the method reveals promising for the whole Rough Indexing Paradigm, we continue developing on compressed streams.

REFERENCES

[1] MPEG-7 Requirements Document V.7: Coding of Moving Pictures and Audio

[2] E. Saez et al., "Global motion estimation algorithm for video segmentation", *Proc. SPIE, VCIP'03*, pp. 1540-1550

[3] R. Ewerth et al. "Estimation of arbitrary camera motion in {MPEG} videos", *Proc. ICPR'04*, pp. 512-515

[4] C. Doulaverakis et al. , "Adaptive Methods for Motion Characterization and Segmentation of MPEG Compressed Frame Sequences", *Proc. ICIAR'04*, pp. 310-317

[5] M. Durik et al, "Robust Motion Characterisation for Video Indexing based on Optical Flow" *Proc. CBMI'01*, pp. 57-64.

[6] P. Bouthemy et al. "A unified approach to shot change detection and camera motion characterization", IEEE Trans. on CSVT, 9(7), pp. 1030-1044