Published in: Proc. IEEE ICIP, International Conference on Image Processing, October 24-27, 2004, Singapore, pp. 385-388.

TEMPORAL VIDEO SEGMENTATION USING GLOBAL MOTION ESTIMATION AND DISCRETE CURVE EVOLUTION

Siripong Treetasanatavorn[1], *Jörg Heuer*[1], *Uwe Rauschenbach*[1], *Klaus Illgner*[1], *and André Kaup*[2]

[1] Siemens AG, Corporate Technology, Information and Communications CT IC 2, Otto-Hahn-Ring 6, D-81739 Munich, Germany
siripong.treetasanatavorn@mchp.siemens.de, {joerg.heuer, uwe.rauschenbach, klaus.illgner}@siemens.com
Tel.: +49 89 636-48409; Fax: +49 89 636-51115

[2] University of Erlangen-Nuremberg, Chair of Multimedia Communications and Signal Processing
Cauerstraße 7, D-91058 Erlangen, Germany; kaup@LNT.de
Tel.: +49 9131 85-27101; Fax: +49 9131 85-28849

ABSTRACT

The identification of syntactic or semantic temporal segments is an important process of video-content analysis. This paper proposes a temporal video segmentation method based on global motion in order to analyze meaningful temporal substructure of camera shots. To ensure that the detected segment optimally contributes to the shot global characteristic, the proposed method exploits a state-of-the-art discrete curve evolution. This technique leads to a subdivision of the global motion trajectory, where each segment of the subdivision has a constant relevant global motion. Experimental results based on standard test sequences acknowledge the method functionality especially for the shots characterized by pronounced camera motion.

1. INTRODUCTION

Temporal video segmentation is regarded as a fundamental and important step of video-content analysis across diverse domains, contexts, and applications [1]. Focusing on the domain of media adaptation, temporal segments and other structuring indices provide useful information for the video summarization preview and flexible presentation preparation [2]. This process is required when a complete video replay is not possible nor preferred due to limited computing resources at the recipient. In the literature, most temporal video segmentation methods [3] adhere to the shot-scene-sequence paradigm [4]. For example, the video Table-of-Content [5] abstracts videos in a similar manner to that of books. Regardless of this paradigm, there exist a number of recent contributions in temporal video segmentation *within a shot* using specific video-content semantics such as motion, e.g. based on object motion [6] and global motion [2].

This paper proposes a new method for temporal segmentation within a camera shot as an alternative to the initial contribution by the authors of this paper in [2]. Under the similar assumption, global motion based on camera motion carries important semantics applicable in temporally partitioning video shots into subshot segments. This segmentation is based on an analysis of the global motion trajectory. In addition, this paper aims that each subshot segment optimally contributes to the global characteristics of a given shot. To achieve this aim, the method exploits the discrete curve evolution technique [7, 8] that efficiently extracts representative parts of the global motion trajectory using the criterion of global contribution.

The paper is organized as follows. Section 2 discusses the global motion estimation used in this paper. Section 3 presents the discrete curve evolution technique and its temporal subshot segmentation method. Section 4 reports the experimental results. The paper is concluded in Section 5.

2. GLOBAL MOTION ESTIMATION

2.1. Estimation of Four-Parameter Global Motion

The global motion model proposed in [9] is adopted and summarized in this section. The model describes the global motion, an estimate of camera motion, in terms of the horizontal and vertical translation t_x and t_y, zooming factor C_F, and rotational velocity φ_z perpendicular to the image plane. Under the condition that φ_z is small, global motion can be modelled by:

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} \approx \begin{pmatrix} C_F - 1 & -C_F\varphi_z \\ C_F\varphi_z & C_F - 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}.$$

This model assumes that the captured scene is flat, i.e., depth to the imaged plane along the optical axis approaches infinity, and the rotational velocities in both axes parallelling to the image plane are small. The four parameters are estimated

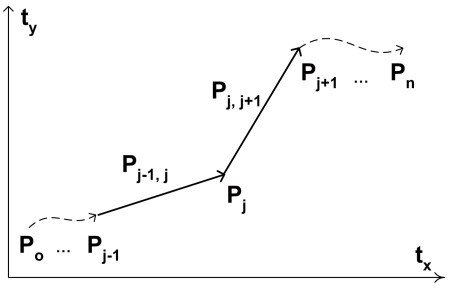

Fig. 1. Global Motion Trajectory Projected onto $t_x - t_y$ Plane

from the displacement vector field $\{(\dot{x}_i, \dot{y}_i)\}$ of coded videos by minimizing the cost function $f(R_1, R_2, t_x, t_y)$:

$$\sum_{i \in V_{bg}} \left[\left(\dot{x}_i - R_1 x_i + R_2 y_i - t_x\right)^2 + \left(\dot{y}_i - R_2 x_i - R_1 y_i - t_y\right)^2 \right]$$

that sums the square errors between the coded displacement vectors on the macroblocks and the corresponding estimates determined from the motion model. Here, we denote $R_1 \hat{=} C_F - 1$ and $R_2 \hat{=} C_F \varphi_z$, and index $i \in V_{bg}$ the coordinates of the displacement vector positions, which contribute to the global motion. The horizontal and vertical displacement vector elements $\dot{x}_i$ and $\dot{y}_i$ correspond to the block position (x_i, y_i).

The parameters R_1, R_2, t_x, and t_y are estimated under the criteria that the partial derivatives with respect to the four parameters of the defined cost function equal to zero, yielding the resulting four global motion parameters $(t_x, t_y, C_F, \varphi_z)$. Only displacement vector fields from P-frames in the analyzed shots are considered in the estimation to reduce the computational complexity.

2.2. Construction of Global Motion Trajectory

The global motion trajectory serves as an analysis basis of the discrete curve evolution (Section 3). A trajectory originates at a vertex $\mathbf{P}_0 = (O_x, O_y, \tau_0)$. Time $\tau_0 = 0$ is referred to the start of the camera shot boundary and constants O_x and O_y are arbitrarily chosen. Terms t_{x,τ_j} and t_{y,τ_j} denote the horizontal and vertical translational estimates of frame j, $j = 1, \ldots, n$ at time τ_j offset to τ_0, respectively. Given the video frame rate F_r, the trajectory comprises a series of edge vectors $\mathbf{P}_{j-1,j}$ $j = 1, \ldots, n$ that connect vertices $\mathbf{P}_{j-1}$ and $\mathbf{P}_j$. The estimation of global motion is only at P-frames; each edge vector $\mathbf{P}_{j-1,j}$ is therefore scaled up with a number of frames $(\tau_j - \tau_{j-1}) \cdot F_r$. Terms $p_{x,j}$ and $p_{y,j}$ of vertices $\mathbf{P}_j$ $(p_{x,j},\ p_{y,j},\ \tau_j)$ are recursively defined as:

$$\begin{cases} p_{x,j} = p_{x,j-1} + t_{x,\tau_j} \cdot (\tau_j - \tau_{j-1}) \cdot F_r \\ p_{y,j} = p_{y,j-1} + t_{y,\tau_j} \cdot (\tau_j - \tau_{j-1}) \cdot F_r \end{cases}$$

Fig. 1 depicts a graphical representation of the global motion trajectory on the $t_x - t_y$ plane.

Discrete Curve Evolution Algorithm (D_n)
$k = n$;
Do
1. find $\mathbf{P}_{j-1,j}$, $\mathbf{P}_{j,j+1}$ in D_k that $R_{j-1,j,j+1}$ is minimal;
2. set $R_{k,min} = \min_{1<j<k-1}(R_{j-1,j,j+1})$;
3. replace $\mathbf{P}_{j-1,j}$, $\mathbf{P}_{j,j+1}$ with $\mathbf{P}_{j-1,j+1}$;
4. shift vertex indices $\mathbf{P}_l = \mathbf{P}_{l+1},\ l = j, \ldots, k-1$;
5. set $k = k - 1$;
Until $R_{k,min} > R_{arm}$ or $k < 2$

Table 1. Discrete Curve Evolution Algorithm

3. TEMPORAL SUBSHOT SEGMENTATION

3.1. Discrete Curve Evolution Algorithm

Let a discrete curve C be initially defined in terms of decomposition D_n of n edge vectors $\mathbf{P}_{j-1,j}$ that connect vertices $\mathbf{P}_{j-1}$ and $\mathbf{P}_j$, $j = 1, \ldots, n$. Let k be the number of segments hierarchically analyzed by the algorithm that sets $k = n$ at the beginning of the evolution. The contribution level of edge pair $\mathbf{P}_{j-1,j}$ and $\mathbf{P}_{j,j+1}$, $j = 1, \ldots, k-1$ to the shape of curve C is gauged by a scalar *relevance measure* [7], $R_{j-1,j,j+1} \hat{=} R\,(\mathbf{P}_{j-1,j}, \mathbf{P}_{j,j+1})$, expressed as:

$$R_{j-1,j,j+1} = \frac{\alpha_{j-1,\ j,\ j+1} \cdot \| \mathbf{P}_{j-1,j} \| \cdot \| \mathbf{P}_{j,j+1} \|}{\| \mathbf{P}_{j-1,j} \| + \| \mathbf{P}_{j,j+1} \|}, \quad (1)$$

given the edge vector length $\| \mathbf{P}_{j-1,j} \|$ calculated by:

$$\sqrt{(p_{x,j-1} - p_{x,j})^2 + (p_{y,j-1} - p_{y,j})^2 + (\tau_{j-1} - \tau_j)^2},$$

and the angle between edge vector pairs, $\alpha_{j-1,\ j,\ j+1} \hat{=} \angle(\mathbf{P}_{j-1,j},\ \mathbf{P}_{j,j+1}) \in [0,\ 180°]$:

$$\alpha_{j-1,\ j,\ j+1} = \arccos\left[\frac{(\mathbf{P}_j - \mathbf{P}_{j-1})^T \cdot (\mathbf{P}_{j+1} - \mathbf{P}_j)}{\|\mathbf{P}_j - \mathbf{P}_{j-1}\| \cdot \|\mathbf{P}_{j+1} - \mathbf{P}_j\|} \right].$$

The higher the *relevance measure* is, the more the edge vector pair contributes to the curve shape. Based on this function, decomposition D_k is evolved (or simplified) to D_{k-1}, $k \geq 2$ by replacing the edge vector pair $\mathbf{P}_{j-1,j}$ and $\mathbf{P}_{j,j+1}$ with a new edge vector $\mathbf{P}_{j-1,j+1}$. The replacement takes place at the pair contributing the least to the global curve structure. For each evolution the number of the edge vectors and vertices shrink by one. The *relevance measure* of this edge vector pair to be substituted is termed minimal *relevance measure* $R_{k,min}$ of decomposition D_k. The evolution carries on until $k < 2$ or the measure $R_{k,min}$ exceeds a predefined *admissible relevance measure*, R_{arm}. The algorithm is summarized in Table 1.

A key property of the algorithm is the order of the edge vector substitutions guided by the search for $R_{k,min}$ in decomposition D_k. If the order is correctly identified, the

evolution leads to a hierarchical structure, that is, the more the evolution iterates, the more significant or representative parts, i.e., edge vectors $\mathbf{P}_{j-1,j}$, $j = 1, \ldots, k$, to the shape of curve C are present in the evolved decomposition D_k.

3.2. Temporal Segmentation by Curve Evolution

Since we assume that the global motion trajectory carries semantic intepretation of a shot, the trajectory evolution is conducted to identify subshot segments that hold significant and representative semantics. For each decomposition D_k, the *relevance measures* of the trajectory edge vector pairs are calculated according to (1). The evolution termination is controlled by R_{arm} (cf. Table 1), which can be expressed by an admissible angle α_{arm} at an admissible length L_{arm}. Term α_{arm} defines the maximal angle size between two connected edge vectors permitting a continuation of the iterative evolution, while L_{arm} defines that of the edge vector length. Calculating R_{arm} using (1) yields:

$$R_{arm} = \frac{\alpha_{arm} \cdot L_{arm} \cdot L_{arm}}{L_{arm} + L_{arm}} = \frac{\alpha_{arm}\, L_{arm}}{2}, \quad (2)$$

where all terms are defined *per-frame*. Term L_{arm} is intuitively parameterized by:

- a proportion T_{prop} of the horizontal and vertical image dimensions C_x, C_y, which implicitly determines a critical level of global translation, and
- a critical time elapse, T_τ.

Given $\overline{\Delta\tau}$, the average time interval between two consecutive vertices, L_{arm} is expressed in a per-frame unit as:

$$L_{arm} = \frac{1}{\overline{\Delta\tau} \cdot F_r} \sqrt{(T_{prop} C_x)^2 + (T_{prop} C_y)^2 + T_\tau^2}.$$

Because $(\frac{T_\tau}{\overline{\Delta\tau} \cdot F_r})^2$ is typically small compared to the first two terms in the square root, it is not taken into consideration. Given $T_{prop}^{\Delta\tau} \hat{=} \frac{T_{prop}}{\overline{\Delta\tau}}$, a substitution of L_{arm} to (2) yields:

$$R_{arm} \approx \frac{1}{2} \cdot \alpha_{arm} \cdot \frac{1}{F_r} \sqrt{(T_{prop}^{\Delta\tau} C_x)^2 + (T_{prop}^{\Delta\tau} C_y)^2}$$

$$= \underbrace{\alpha_{arm} T_{prop}^{\Delta\tau}}_{configurable} \underbrace{\frac{\sqrt{C_x^2 + C_y^2}}{2F_r}}_{known\ a\ priori}. \quad (3)$$

Therefore, the *relevance measure* R_{arm} is set by three *a priori* parameters C_x, C_y, F_r and two configurable parameters, α_{arm} and $T_{prop}^{\Delta\tau}$.

At the last decomposition D_k, k edge vectors $\mathbf{P}_{j-1,\,j}$, $j = 1, \ldots, k$, are defined as subshot segments. Each segment is determined by two vertices $\mathbf{P}_{j-1}$, $\mathbf{P}_j$ at time τ_{j-1} and τ_j, respectively.

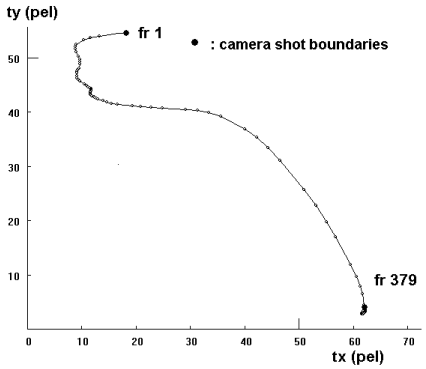

Fig. 2. Global Motion Trajectory from Sequence *Foreman*

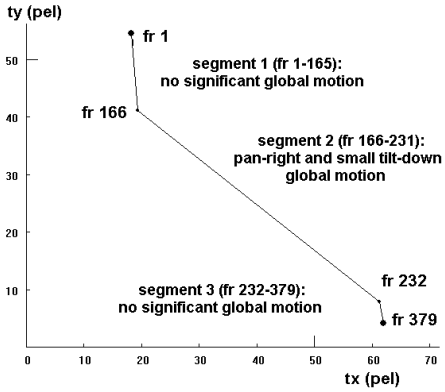

Fig. 3. Temporal Subshot Segmentation Result from Sequence *Foreman* Using $T_{prop}^{\Delta\tau}$=0.8/sec, α_{arm}=3°/fr; Last $D_k,\ k = 3$

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

The experiment aimed at evaluating the functionality and efficiency of the proposed method. The implementation of temporal segmentation was on a platform comprising a compressed video (MPEG-1) parser, spatial segmentation and key-frame selection tool [2]. To set R_{arm} in (3), α_{arm} was intuitively chosen at 3°/fr. For instance, at $F_r = 25$ fps, 75°/sec is a critical angle change in camera translation. $T_{prop}^{\Delta\tau}$ was chosen at a sufficiently large value such that the insignificant or spurious camera motion was disregarded by the evolution. Whereas, $T_{prop}^{\Delta\tau}$ should be configured above the lower bound of the expected critical translation. This setting enables the trajectory vertices influenced by the intended camera translation to remain until the last decomposition. These are two criteria used for setting $T_{prop}^{\Delta\tau}$. Empirically, $T_{prop}^{\Delta\tau}$ was set at 0.8/sec or the per-second camera translation at 80% of the image dimensions was critical in declaring subshot segments. For example, Fig. 2 depicts the global motion trajectory constructed from the sequence *Foreman*. Fig. 3 depicts its last decomposition that defined

Group	Exp	CD	MD	FA	Recall	Precision
A	37	32	5	1	0.86	0.97
B	63	48	15	9	0.76	0.84
Total	100	80	20	10	**0.80**	**0.89**

Table 2. Summary of Temporal Subshot Segmentation Results from Two Shot Groups Using $T_{prop}^{\Delta\tau}$= 0.8/sec, α_{arm}= 3°/fr

the subshot segments. F_r, C_x and C_y were set at 24 fps, 176 and 144 pel, respectively.

In the evaluation process, two groups of 31 test video shots were selected from two well-known sequences *Foreman* (*FM*) and *Stefan* (*ST*), as well as parts of three standard test sequences *Documentary about buildings, Lancaster Television* (*LA*), *Documentary about a village "Santillana del Mar" RTVE* (*VL*), and *Edited home video LGERCA* (*LC*) from [10]. Shots from sequences *ST*, *LA* and *VL* were categorized to test group A. This category contained mainly smooth camera motion with minimal distortion. On the contrary, the camera motion from shots in group B (from *FM* and *LC*) was often ambiguous and interfered by irregular random camera movement. 100 ground-truth subshot segments were manually selected from the 31 test video shots. Unlike most evaluation methods found in the literature, the ground-truth segments in this paper were attributed by both *the characteristic of the pronounced camera motion* (e.g. pan-right, tilt-up, etc.) and *their approximate time intervals*. This is because the human anticipation of subshot segment boundaries characterized by coherent camera motion is highly subjective and difficult to be evaluated. Based on this ground truth, performance of the detected segments was measured in terms of the *Recall* ($\frac{CD}{CD+MD}$) and *Precision* ($\frac{CD}{CD+FA}$) *rates*, where CD, MD, and FA denote the number of the correct, missed and false-alarmed detection, respectively. The results are summarized in Table 2.

From shots in group A, the recall and precision rates were relatively high (0.86 and 0.97, respectively). Such results were originated from the reliable camera motion estimates and trajectories which established a stable analysis foundation. On the other hand, the shots from group B showed unclear structures which increased the amount of missed detection, thereby causing an inferior recall rate (0.76). However, a good precision rate performance was achieved (0.84) from shots in this group because the method is capable of differentiating the intended camera motion from the spurious ones. This action was carried out by investgating the relative influence of each edge vector pair in terms of the *relevance measure* throughout the shot. This property is an advantage of the method.

In terms of the complexity, the method requires in each evolution to find an edge vector pair of the smallest *relevance measure* in order to guarantee the right order of the edge vector substitutions. Given a number of vertices n at the initial global motion trajectory, the algorithm complexity has an order of $O(n^2)$, if the number of edge vectors k at the last decomposition D_k is much smaller than n.

5. CONCLUSIONS

This paper proposes a temporal video segmentation method using global motion estimation and discrete curve evolution. The method hierarchically extracts significant structures of the global motion trajectory and interprets parts of the resulting evolved trajectory as subshot segments. Experimental results acknowledge that the technique is suitable for abstracting video shots, especially when global motion in the shots is pronounced.

6. ACKNOWLEDGMENT

The authors would like to thank Prof. Dr. Ulrich Eckhardt from Universität Hamburg, Institut für Angewandte Mathematik, for a lot of inspiring discussions.

7. REFERENCES

[1] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of video-content analysis and retrieval," *IEEE Multimedia*, vol. 9, no. 3, Jul.-Sep. 2002.

[2] A. Kaup *et al.*, "Video analysis for universal multimedia messaging," in *Proc. IEEE SSIAI*, Apr. 2002, pp. 211–215.

[3] I. Koprinska and S. Carrato, "Temporal video segmentation: A survey," *EURASIP Sig. Proc.: Image Communication*, vol. 16, no. 5, pp. 477–500, Jan. 2001.

[4] J. Monaco, *How to Read a Film: The World of Movies, Media and Multimedia*, chapter The Language of Film: Signs and Syntax, pp. 151–225, Oxford University Press, 2000.

[5] X.S. Zhou *et al.*, *Exploration of Visual Data*, chapter Constructing Table-of-Content for Videos, pp. 53–73, Kluwer Academic Publishers, 2003.

[6] Y. Fu *et al.*, "Temporal segmentation of video objects for hierarchical object-based motion description," *IEEE Trans. Image Processing*, vol. 11, no. 2, pp. 135–145, Feb. 2002.

[7] L.J. Latecki and R. Lak¨amper, "Convexity rule for shape decomposition based on discrete contour evolution," *Computer Vision and Image Understanding (CVIU)*, vol. 73, no. 3, pp. 441–454, Mar. 1999.

[8] R. Lak¨amper, *Formbasierte Identifikation zweidimensionaler Objekte*, Ph.D. thesis, Fachbereich Mathematik, Universit¨at Hamburg, Germany, 2000.

[9] J. Heuer and A. Kaup, "Global motion estimation in image sequences using robust motion vector fi eld segmentation," in *Proc. ACM Multimedia*, Nov. 1999, pp. 261–264.

[10] MPEG, "Licensing agreement for the MPEG-7 content set," ISO/IEC JTC1/SC29/WG11/N2466, Atlantic City, 1998.