Subject Category: image processing, image motion analysis, motion estimation, image processing, image sequences

Retrieve video clips using the global motion information

Tianli Yu and Yujin Zhang

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

In this paper, a new scheme is proposed to extract the global motion model from general video sequence. Since global camera motion often conveys the semantic meaning of the video's content, these global model parameters are used as features for video retrieval. Experimental results show our scheme is quite effective in locating video segments with similar motion content in football game sequences.

Introduction: In general video sequences, the movement of the camera always follows the place where the important activity goes on. This means the camera motion contains cues of the semantic meaning of the video, especially in some constrained domain. One good example of this is the football game. In football game video, the camera's movement is often focusing on the trajectory of the football and the position of the players. This information could be used as the feature for the retrieval of the football sequence. In football games, color and texture feature based retrieval methods will generally failed since the scene won't change too much from the point of view of color and/or texture. Cameral motion based feature could be a good alternative. The advantage of using global camera motion feature is that it could be extracted easily and robustly. In this paper we propose a simple method to estimate the bilinear motion model parameter of the global camera motion and use the extracted global motion model as the feature for the retrieval of video clips.

Global Motion Estimation: The motion of pixels in two consecutive video frames can be divided into global motion and local motion. Global motion is caused by the movement of camera and could be calculated by 7estimating several model parameters using the linear estimation methods [1]. We choose the bilinear motion model and least square estimation to estimate the global camera motion. According to the bilinear model, the global motion vector (u,v) of a point (x,y) could be expressed as:

$$\begin{cases} u = a_0 + a_1 x + a_2 y + a_3 xy \\ v = a_4 + a_5 x + a_6 u + a_7 xy \end{cases} \qquad (1)$$

where $a_0 ... a_7$ are the model parameters.

Theoretically, with a set of more than 4 different point's motion vectors, we could estimate the 8 bilinear motion parameters using the least square estimation method. We divide the video frame into blocks and use block match algorithm to get these motion vectors.

In block match, we could choose relatively large block size to reduce the noise caused by local object motion and luminance variation. Even though, there will be some mismatched block that will give erroneous motion vectors, especially in low texture area. Figure 1 shows a block match result of two consecutive frames in a football game video. In order to eliminate the interference of the erroneous block match results to the global motion model, we use an iterative rejection [2] algorithm to make the estimation of the global motion model more robust. Our algorithm is summarized as follows:

1. Mark all the blocks as "in-rejected"
2. Use all the "in-rejected" block data to estimate the bilinear motion model parameter, with the least square estimation method, to minimize the sum of square estimation error

$$E_{all} = \sum_{i=1}^{N} (u_i - \hat{u}_i)^2 + (v_i - \hat{v}_i)^2 \qquad (2)$$

where $(\hat{u}_i, \hat{v}_i)$ are motion vectors calculated from equation (1) using the estimated parameters, (u_i, v_i) is the block match result of the i_{th} block.

3. Calculate the average estimation error of the model parameters,

$$E_{avg} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(u_i - \hat{u}_i)^2 + (v_i - \hat{v}_i)^2} \qquad (3)$$

Use E_{avg} as a threshold, mark the blocks whose estimation error is larger than the E_{avg} as “rejected”.

4. If the “rejected” blocks is different from those rejected at last iteration, then go to step 2, else stop the iteration.

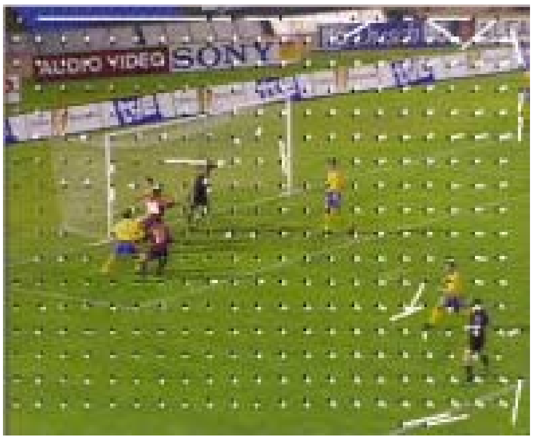

Figure 1. Block match result of two consecutive frame in a football game sequence.

When the iterative rejection algorithm terminate, we could eliminate most of the erroneous blocks and get the global model parameters between two consecutive frames accurately, figure 2 shows the reconstructed global motion vector from the estimated bilinear model, the gray blocks are those marked as “rejected” in the last iteration.

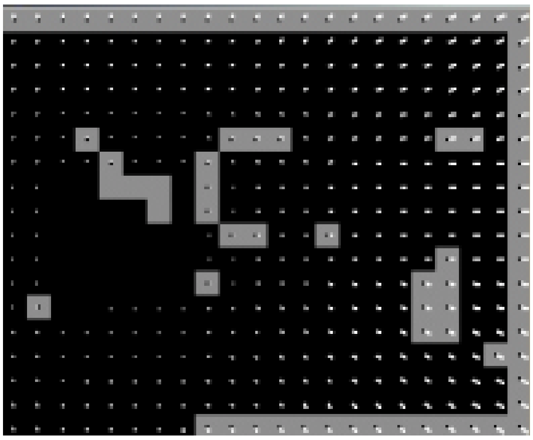

Figure 2. Reconstructed global motion vector from the estimated bilinear motion model, (gray block is the “rejected” blocks in the last iteration)

Distance measurement: To find the video clips with similar global motion, we need to define a measurement to calculate the distance between two video clips. First, we define the distance between two different global motion models. The distance between two global-motion models m_1 and m_2 is the sum of the square of motion vector difference reconstructed from their motion models, as expressed in equation (4)

$$D(m_1,m_2)=\sum_{x,y}[u_{m1}(x,y)-u_{m2}(x,y)]^2+[v_{m1}(x,y)-v_{m2}(x,y)]^2 \tag{4}$$

Assume m_1's model parameters are $a_{10}...a_{17}$ and m_2's model parameters are $a_{20}...a_{27}$, we could write equation (4) as

$$\begin{aligned}D(m_1,m_2)= &\sum_y\sum_x[(a_{10}-a_{20})+(a_{11}-a_{21})x+(a_{12}-a_{22})y+(a_{13}-a_{23})xy]^2\\ &+[(a_{14}-a_{24})+(a_{15}-a_{25})x+(a_{16}-a_{26})y+(a_{17}-a_{27})xy]^2\end{aligned} \tag{5}$$

Expand and simplify equation (5), we could get

$$\begin{aligned}D(m_1,m_2)= &\left[(a_{13}-a_{23})^2+(a_{17}-a_{27})^2\right]\sum_x x^2\sum_y y^2\\
&+\left[(a_{11}-a_{21})^2+(a_{15}-a_{25})^2\right]\sum_x x^2\sum_y 1\\
&+\left[(a_{12}-a_{22})^2+(a_{16}-a_{26})^2\right]\sum_x 1\sum_y y^2\\
&+\left[(a_{10}-a_{20})^2+(a_{14}-a_{24})^2\right]\sum_x 1\sum_y 1\\
&+2\left[(a_{13}-a_{23})(a_{11}-a_{21})+(a_{17}-a_{27})(a_{15}-a_{25})\right]\sum_x x^2\sum_y y\\
&+2\left[(a_{13}-a_{23})(a_{12}-a_{22})+(a_{17}-a_{27})(a_{16}-a_{26})\right]\sum_x x\sum_y y^2\\
&+2\left[(a_{13}-a_{23})(a_{10}-a_{20})+(a_{17}-a_{27})(a_{14}-a_{24})\right]\sum_x x\sum_y y\\
&+2\left[(a_{11}-a_{21})(a_{12}-a_{22})+(a_{15}-a_{25})(a_{16}-a_{26})\right]\sum_x x\sum_y y\\
&+2\left[(a_{12}-a_{22})(a_{10}-a_{20})+(a_{16}-a_{26})(a_{14}-a_{24})\right]\sum_x 1\sum_y y\\
&+2\left[(a_{11}-a_{21})(a_{10}-a_{20})+(a_{15}-a_{25})(a_{14}-a_{24})\right]\sum_x x\sum_y 1\end{aligned} \tag{6}$$

If the width of the frame is M and the height of the frame is N, then we have

$$\sum_x 1=M\ ,\ \ \sum_y 1=N \tag{7}$$

$$\sum_x x=\frac{M(M+1)}{2}\ ,\ \ \sum_y y=\frac{N(N+1)}{2} \tag{8}$$

$$\sum_x x^2=\frac{M(M+1)(2M+1)}{6}\ ,\ \ \sum_y y^2=\frac{N(N+1)(2N+1)}{6} \tag{9}$$

Thus, with the parameters of the two models as well as the width and height of the frame, we could calculate the distance between these two models using equation (6).

For a video clip we could estimate the global motion model for each consecutive frames. In this way we got a sequence of global motion model. To compare the distance between 2 equal length video clips, we define the distance between these two clips as the sum of distance of their corresponding global motion models in the sequence, as shown in equation (10)

$$D(V_1,V_2)=\sum_{i=1}^{L}D(m_1(i),m_2(i)) \tag{10}$$

Query by Example: Given a small example of video clip V_q, we want to find the video segments with similar global motion in a long video sequence V_0. Assume V_q has length of L, first we take a L length clip V_s

from V_0 at starting time t and compute the distance between V_q and V_s with equation (10). By shifting the starting time t from one frame to another, we could get a function $D_{Vq,V0}(t)$ showing the relation of distance and the segment starting time, as defined in equation (11)

$$D_{Vq,V0}(t) = \sum_{i=1}^{L} D(m_q(i), m_0(i+t)) \tag{11}$$

The local minimum points of $D_{Vq,V0}(t)$ are the possible query results, we sort all the local minimum points of $D_{Vq,V0}(t)$ according to their distance and those with smallest distances are our final query results.

Experimental Results: Our experimental sequence comes from a 14 min football video, first we use our global motion model estimation algorithm to get the model sequence and use query by example method to find the clips with similar global motion. In the video sequence we manually locate 6 ground-truth video segments with the camera doing an operation of left pan along with a zoom-in, each with 25 frames in length. Usually in football video, a zoom-in indicates a sudden event that draws most of the attention, such as a shoot to the goal; the left pan means the event is moving from right to left. We use each of the 6 segments as the query example and positions of the correctly retrieved segments in the top 10 retrieved segments are summarized in Table 1:

Table 1 Successfully retrieved video clips' position

	Correctly retrieved clip's position in the top 10 result				
Query Sequence	1st	2nd	3rd	4th	5th
Z1	1	2	3	4	--
Z2	4	6	--	--	--
Z3	1	2	4	10	--
Z4	1	2	3	5	--
Z5	1	2	3	5	8
Z6	1	3	--	--	--

In Table 1, the row is the query result retrieved by one of the 6 ground-truth segments, the column shows the index of the *i*th correctly retrieved video segments in the top 10 retrieved results. In the results, though there are still some clips missing, our scheme can return many of the correct clips in the first round retrieval results.

Conclusions: A scheme to retrieve the video clips using its global motion information is described. A simple method to estimate the global motion model was proposed and the distance measurement was defined to calculate the similarity of two video clips. Query by example results show this system could effectively be used to retrieve clips with similar camera motion. In some constrained domain video, such as football game, this also means video clips with similar semantic meaning could be retrieved by our scheme.

Reference:

1. Eung Tae Kim, Hyung-Myung Kim, "Efficient linear three-dimensional camera motion estimation method with applications to video coding", Opt. Eng. 1998, 37(3), pp1065-1077
2. Tianli Yu, Yujin Zhang, "Motion feature extraction for content-based video sequence retrieval". in Internet Imaging II, SPIE 4311, 2001, pp.378-388