Key Frame Selection from MPEG Video Data

Omer N. Gerek[1] and Yucel Altunbasak[2]

[1]Bilkent University, Dept. of Electrical and Electronics Engineering,
Bilkent, Ankara TR-06533, Turkey
E-mail: gerek@ee.bilkent.edu.tr
Phone: (90) 312-266 4307 Fax: (90) 312-266 4126

[2] Department of Electrical Engineering and Center for Electronic Imaging Systems
University of Rochester, Rochester, NY 14627
E-mail: altunbas@ee.rochester.edu

Keywords : Key Frame Selection, Video Indexing, Video Database, MPEG

Abstract

This paper describes a method for selecting key frames by using a number of parameters extracted from the MPEG video stream. The parameters are directly extracted from the compressed video stream without decompression. A combination of these parameters are then used in a rule based decision system. The computational complexity for extracting the parameters and for key frame decision rule is very small. As a result, the overall operation is very quickly performed and this makes our algorithm handy for practical purposes. The experimental results show that this method can select the distinctive frames of video streams successfully.

1 INTRODUCTION

A good video database management system requires efficient ways of abstracting the video information. Identification of *key frames* in a video stream is an important task for summarizing the content of a video.[1] Since most of the video data are already in compressed format, it is usually desired to perform this operation over the readily produced compressed video data. The most commonly used compression methods for video data are MPEG-I or MPEG-II. During key frame detection, one should avoid full decompression of the data in order to reduce the computational complexity.[2,3] In the following sections, we describe how we perform key frame detection in different video contents. Different frame types in an MPEG video stream are examined using different algorithms, and a special emphasis is given to the situations of panning and zooming.

Yucel Altunbasak is now affiliated with Hewlett-Packard Laboratories,Palo Alto, CA.

SPIE Vol. 3024 • 0277-786X/97/$10.00

2 METHOD

2.1 Sharp Scene Cut Detection

In regular video streams, the sharp scene cuts are relatively easy to locate, so they should be quickly determined. To speed up the sharp scene cut detection, we first bound the sequence interval in which the scene cut appears by examining only I frames. I frames are the frames that are intra coded and do not require data from the other frames in a sequence. Typically, an MPEG stream consists of several I frames, and P and B frames, which are predicted frames, between every two I frames. In this way, the sequence segment between two I frames is marked as "No scene cut" or "Scene cut" depending on some criteria. If no sharp scene break is found, then the algorithm proceeds to the next I frame interval.

We apply two tests to determine whether a scene cut occurs in an I frame interval. The first test utilizes the histogram content of the frame and is also used by other researchers.[1] The histogram of the DC coefficients of each macro block is constructed for each I frame, then the histogram difference between two I frames is defined as

$$HD(I_i, I_{i+1}) = \left[\frac{\sum_{k=1}^{N}(H_i(k) - H_{i+1}(k))}{\sum_{k=1}^{N}(H_i(k) - H_{i+1}(k))}\right]^2, \quad (1)$$

where I_i, and I_{i+1} are the i^{th} and $i+1^{th}$ intra coded frames, respectively. H_i, and H_{i+1} is the DC histograms of i^{th} and $i+1^{th}$'s I frames. N is the number of bins in the DC histogram.

If this $HD(I_i, I_{i+1})$ value is greater than a threshold δ_1, the video segment (I frame interval) between frame I_i and I_{i+1} is marked as "Potential Scene Cut". The value of δ_1 proves to be important for this method to properly mark the intervals containing sharp scene cuts. This parameter is determined based on the statistics of the $HD(I_i, I_{i+1})$ values. More specifically, the mean μ_k and standard deviation σ_k of $HD(I_i, I_{i+1}), i = 1...k$ is calculated, then δ_1 is set to be $\delta_1 = \mu_k + 3\sigma_k$.

If $HD(I_k, I_{k+1})$ is less than δ_1, then the algorithm proceeds to the next I frame interval. If, however, histogram difference is greater than the current histogram threshold, the current I frame interval either contains a sharp scene cut or a significant amount of object motion. It is important to note that unlike the pixel-based histogram, macro-block based histogram can not get rid of object motion completely. If objects move multiple of 16 pixels, which is the horizontal or vertical size of a macro-block, macro-block based histogram difference would not have any term due to object motion. But, in reality this is usually not the case. Fortunately, the slowly varying intensity patterns of objects in a natural scene allows DC histogram perform remarkably well regardless of the amount of motion. However, in some cases, it may still be necessary to test for "object motion" before concluding a "sharp scene cut".

The direct approaches for such a determination, such as (object) motion estimation, is too computationally expensive. We propose to use low-pass filtered DC image to construct DC histogram, where a DC image refers to an image formed by DC coefficients of the transform coefficients for each macroblock. The low-pass filtered histogram $LHD(I_k, I_{k+1})$ is compared against another threshold, δ_2, in this second round of test. If this test yields a "Scene Cut", the interval is said to be containing a "Sharp Scene Cut".

Once the I frame interval is identified, we can then search for this interval to determine precisely where the break occurs by examining I, P and B type of frames within this video segment.

The search algorithm is different for each type of frame, and is summarized as follows:

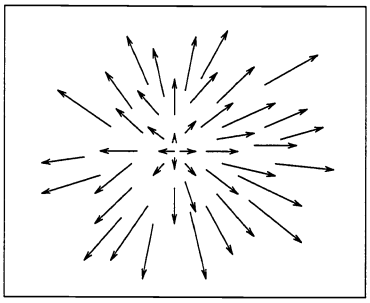

Figure 1: Motion vectors of Ideal Zoom

2.1.1 P frames

If an abrupt scene change occurs just before a P type frame, motion estimation should not have been of help in encoding. Therefore, many of the macroblocks should have been coded as "Intra" which means that a proper motion could not be found and the block is coded without a prediction. If the ratio of macroblocks coded in "Intra" mode is higher than the threshold δ_p, it is appropriate to select this frame as a "Scene Break Boundary".

The selection of the threshold δ_p is again important. In this work, we specify this threshold by analyzing the statistics of intra coded macroblock ratio over the video sequence. This statistics can also be formed adaptively during the key frame selection process.

2.1.2 B frames

If the current frame is of B type and it is a new scene, almost all the motion vectors should be of unidirectional. This is basically due to the fact that if the current or the next frame is a scene cut, then either forward or backward motion prediction should fail. More specifically, if all motion vectors are forward motion vectors, then the current B frame is a scene cut because the previous frame is significantly irrelevant to the current frame. On the contrary, if all motion vectors are backward, then the next frame is a scene cut due to the same reason. To this respect, a B frame is selected as a key frame if more than 90% of the motion vectors have the same direction.

2.1.3 I frames

The key frame decision of an I type frame solely depends on whether the P or B frames inside the I frame interval are selected as key frames, or not. If the scene cut is not found at any of the P and B frames, it is assumed to be at the first I frame since we know that a scene cut exist in this I frame interval. Note that I frame interval contains the beginning I frame, but excludes the ending I frame. Therefore it only contains a single I frame which will be chosen.

2.2 Zoom Detection

An ideal zoom pattern is depicted in Fig. 1. Fig 2 shows the number of motion vectors for each angle θ, that the motion vectors make with the positive horizontal axis, for this ideal zooming. On the other hand, Fig 3 depicts the number of motion vectors vs. size of motion vectors for the same zoom.

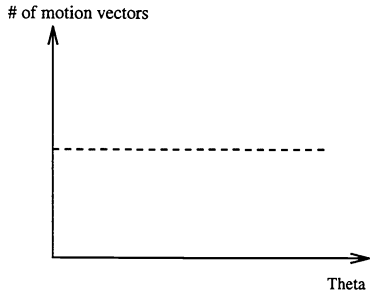

Figure 2: Number of motion vectors versus theta for ideal zoom

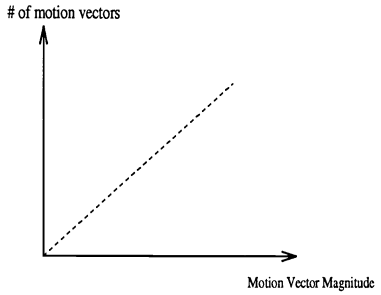

Figure 3: Number of motion vectors versus size of motion vectors for ideal zoom

For each P frame, the motion histograms which will somehow resemble Figs. 2 and 3 are computed. Then,

- Calculate the variance of the number of motion vectors vs. angle θ histogram. If it were to be flat as in Fig. 2, then the variance should have been very high. If the variance is higher than a predefined threshold δ_z, then a zoom is likely to exist. However, we check the number of intra coded macroblock just to make sure that a zoom really occurs because since zooming usually yields occlusion, it yields some intra coded macroblocks. Therefore, even if variance is very high, if the ratio of intra coded block is under a predefined percentage, we will not consider the frame as a "Zoom" frame. If both conditions hold, the frame can be marked as a "Zoom" frame.

If the percentage of "Zoom" marked frames is above %90 in a sequence segment of length at least 25, then this sequence segment is labeled as "Zooming segment".

2.3 Pan Detection

The case of panning is quite similar to the case of zooming. An ideal panning pattern is depicted at the left side of Fig. 4. The right side of Fig 4 shows the number of motion vectors for each angle θ, that the motion vectors makes with the positive horizontal axis for this ideal panning.

For each P frame, the motion histograms which should look like the right sides of Figs. 4 and 5 are computed. Then, perform the following:

- Calculate the variance of the number of motion vectors vs. angle θ histogram. If it were to be like

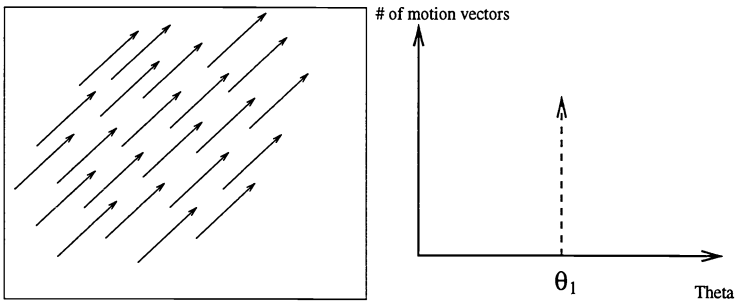

Figure 4: An "ideal pan" and "Theta versus number of motion vectors" for ideal pan.

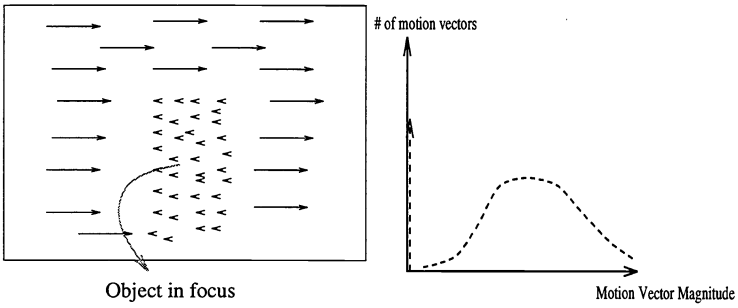

Figure 5: Motion vectors for a typical pan with camera tracking the object, and the theta versus number of motion vectors for this scheme.

an impulsive as in Fig. 4, then the variance would have been very small. If the variance is less than a predefined threshold δ_p, then pan is likely to occur. Mark this frame as "Pan".

If the percentage of "Pan" marked frames is above %90 in a sequence segment of length at least 25, then this sequence segment is labeled as "Panning"

However, the situation depicted above explains "Camera Panning". That is, only camera is panning, no significant object movement is present. However, in many cases camera tracks an object, such as a football player, or a ball, such that the object in focused is hold constant while camera is panning. Therefore, there might be many zero motions in such a case. In this case, the ideal pan histogram would have to be modified as in the right side of Fig.5. In this case, we simply discard all zero motion vectors in variance and histogram calculations.

2.4 Gradual Scene Transitions

Other types of camera motions are considered as "gradual scene transitions" for the time being. The current I frame is compared against the last selected key frame via DC histogram as explained in Section . If the histogram difference between the current frame and last key frame is above a threshold, then the current I frame is selected as a gradual scene transition key frame.

Figure 6: Frame no 90 and 125

3 RESULTS

For appropriately selecting the key frames, the comparison thresholds have to be chosen according to the general characteristics of the sequence. In our experiments, we used CIF size (352×288) frames for the MPEG sequence. The frame rate is 8 frames/sec. The motion measure and the thresholds should be modified accordingly if the frame rate is altered because the motion characteristics of the sequence changes according to the frame rate. The Group Of Picture (GOP) structure has the sequence : IPBBBPBBBPBBBPIPBBB... With the given settings, our algorithm detected the frames indexed by 0 90 125 167 186 238 259 349 397 437 462 472 537 557 588 615 637 672 720 and 800 in the set of 801 frames. The frames 90 and 125 are given in Fig. 6.

4 CONCLUSIONS

We integrated the previous state of the art in temporal video segmentation, and proposed a few key modifications, which increased the performance with the addition of very little computational cost. However, more complex camera operations and special effects such as dissolve, fade in, fade out, and wipe should be modeled and integrated with the proposed system. Current research focuses on this direction.

5 References

[1] Borko Furht, Stephen W. Smoliar, and HongJiang Zhang, "Video and Image Processing in Multimedia Systems," *Kluwer Academic Publishers,* 1995.

[2] Jianhao Meng, Yujen Juan, Shih-Fu Chang, "Scene Change Detection in a MPEG Compressed Video Sequence," *IS&T/SPIE Symposium Proc.*, Vol. 2419, Feb. 1995, San Jose, California

[3] Hain-Ching, H. Liu and Gregory L. Zick, "Scene Decomposition of MPEG Compressed Video," *IS&T/SPIE Symposium Proc.*, Vol. 2419, Feb. 1995, San Jose, California