REPRESENTATION OF MOTION ACTIVITY IN HIERARCHICAL LEVELS FOR VIDEO INDEXING AND FILTERING

Xinding Sun, B. S. Manjunath*, Ajay Divakaran*[+]

*Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106
{xdsun, manj}@ ece.ucsb.edu

[+]Mitsubishi Electric Research Laboratories, 571 Central Avenue, #115 Murray Hill, NJ 07974
ajayd@merl.com

ABSTRACT

A method for video indexing and filtering based on motion activity characteristics in hierarchical levels is proposed. To extract motion activity information, an MPEG (MPEG-1/2) video is first adaptively segmented into hierarchical levels with fixed percentage of original video length based on P-frame macroblock motion information. Three motion activity characteristics-motion intensity which represents the degree of change in motion, motion intensity histogram which represents the temporal statistics of motion intensity, and spatial descriptor which represents the spatial attribute of motion, are then computed to represent different levels of video. The descriptors from different levels are used selectively in different steps of video indexing and filtering. Experimental results show the proposed method is fast and effective, and provides a powerful video indexing and filtering tool.

1. INTRODUCTION

Motion as a as a visual feature has been used in content-based video retrieval systems [1], [3], [5], [10], [11]. Typically, information about the dominant motion in a given region of interest is encoded for search and retrieval. In contrast, this paper proposes a new representation to characterize the spatial and temporal distribution of motion in a given video segment. It describes characteristics scene change in terms of the motion intensity, the corresponding temporal distribution of motion intensity using a motion intensity histogram, and the spatial distribution of motion. Our objective is to test the effectiveness of motion activity descriptors for video indexing and filtering purpose.

The proposed method works with MPEG encoded video and uses the macro-block information in feature extraction. A given video scene is hierarchically partitioned into a number of segments, coarse to fine, with a "fine" segment containing very few frames, and a "coarse" segment containing a large number of frames. In segmenting video into hierarchical levels, the P-frame macroblock information of an MPEG video is used. There are two reasons for using P frame information. First, digital video possesses redundant information; therefore, P frames are good temporal samples of the original video, and have been used in many applications [9]. Second, P frames are encoded with macroblock information that can be processed quickly, as discussed in the next section. We adaptively segment video into levels with fixed percentages (1% to 20%) of original video length using the method proposed in [6]. This is based on the observation that the pattern of motion change can be across the shots, while within a shot the motion pattern can also change significantly. For example, the motion in a sports video shot can vary from motionless during the pause of the game to significant large motion when many players are running. The levels with fixed percentages with respect to the whole video length also provide an effective way of comparing video segment quantitatively.

The motion activity descriptors at different levels can be obtained once the video is segmented. We present a strategy that uses motion intensity histogram for coarse level video segments and uses the spatial descriptor for fine level segments for video indexing and filtering.

2. HIERARCHICAL VIDEO SEGMENTATION

In much of the previous research work, video was segmented into shots based on a variety of features like color, texture, motion, etc. The reason they work well is that all these features usually change sharply at shot boundaries. However, different features generally behave differently in inter-shot and intra-shot changes. Therefore it is of interest to investigate one single feature, such as motion, throughout the video segmentation.

2.1. Motion Intensity

The intensity of motion—the level of motion activity and the change in it—is used for video segmentation. The following discussion applies to only P-frames in an MPEG coded video.

In order to reduce the bit rate, some macroblocks in the MPEG P frames are coded using their differences with the corresponding reference macroblocks. Note that the No_MC macroblocks in MPEG video have no motion compensation [4] The No_MC macroblock can also be categorized into two types, one is intra-coded and the other is inter-coded. In a special case, when the macroblock perfectly matches its reference, it is skipped and not coded at all. To simplify the illustration, the skipped frames are categorized the same as No_MC inter-coded frames. The No_MC inter-coded macroblock has zero motion [8].

According to the definition of inter-coded No_MC, when the content of video changes are not too significant, and thus many macroblocks match their references, the inter-coded No_MC macroblock numbers would be high in a P frame. For

example, pauses in sports games often coincide with small object motion and the corresponding inter-coded No_MC macroblock numbers would be very high. On the other hand, when the content of the video changes rapidly, and thus many macroblocks cannot be matched by their reference frames, the inter-coded No_MC macroblock numbers would be small in a P frame. Here, we define the α-ratio of a P frame as:

$$\alpha = \frac{Number\ of\ inter\ No\_MC\ Macroblocks}{Total\ Number\ of\ Frame\ Macroblocks} \quad (1)$$

From our experiments, we find that this ratio is a good measure of scene motion intensity change and it conforms with human perception very well. The higher the ratio is, the lower the scene motion intensity change is. Figure 1 shows two frames from the same shot in a football video, the first one extracted from the start of a play, which has a high α = 86%, and the second one corresponding to the play in progress, which has a low α = 5%. The two frames are from the same video shot though their motion characteristics are significantly different.

(a) α=0.86 (b) α=0.05

Figure 1.Two frames with different inter-coded No_MC ratios from the same shot.

A logarithmic compandor is used to quantize α-ratio and this non-linear scaling matches subjective motion perception reasonably well [7]. By using this method, we can keep the quantization step higher for high ratio values. The ratio is compressed using the μ-law characteristic:

$$G_u(\alpha) = Q\frac{\log(1+u\alpha/Q)}{\log(1+\mu)}\ ; \quad \alpha \le Q \quad (2)$$

Where the parameter μ is set to 255 and Q is set to 1. $G_u(\alpha)$ is used to represent motion intensity.

2.2. Adaptive segmentation of video into levels

Given a video sequence, we would like to obtain a pre-determined fixed number of representative frames to describe the video. This number can be 1%, 5%, 10%, 20%, etc., of the total number of frames in the original video. These representative frames can be used for video summarization and indexing. [6]. On the other hand, once these representative frames are extracted, their temporal positions in the video basically segment the whole video into smaller units. Correspondingly, the numbers of these units are 1%, 5%, 10%, 20%, etc., of the number of frames in the video. These units represent a video in *hierarchical levels*. The coarser of video correspond to smaller percentage of units and they coarsely represent video content. The finer levels of video get into more details of video content, and correspond to a larger percentage of units. For example, units at 10% (finer) level can be a subset of those units at 5% (coarser) level. The advantage of this segmentation is that we can compare video content change quantitatively in terms of their duration.

Even though not normalized by the MPEG standard, the positions of P frames in a video stream usually take fixed positions in a Group of Pictures (GOP) [4], and consequently fixed positions in a video. Therefore, they are good temporal samples of original video. Thus, instead of processing the video frames one by one, we process only the P-frames for temporal segmentation.

A video segment can be modeled as a trajectory of multidimensional feature points in a multidimensional space. The nature of the spatial distribution of the points that represent their corresponding video frames can be described as clusters connected by abrupt or gradual changes. This nature of the distribution of points provides a sound basis for clustering techniques. Since we are analyzing the trajectories of feature points in temporally localized units of frames, it is possible to use the change in consecutive representative frames to represent the change within a unit. Given a unit U_i and a selected feature P, we define the *unit change* as follows:

$$Change(U_i) = D_c(R_{pi}, R_{p(i+1)}) \quad (3)$$

Where $R_{pi}, R_{p(i+1)}$ are the consecutive frames that bound the unit, $D_c(\bullet)$ computes the difference of two frames with respect to a selected feature. Here the feature we choose is the motion intensity measure. Therefore

$$D_c(R_{pi}, R_{p(i+1)}) = \left|G_u(\alpha_i) - G_u(\alpha_{i+1})\right| \quad (4)$$

From an optimization point of view, given a selected feature, the objective of representative frame extraction is to divide a video into units that have very similar unit changes. This can be formulated as the minimization of:

$$\sum_{i=1}^{N'-2}\sum_{i=j+1}^{N'-1}\left|Change(U_i) - Change(U_j)\right| \quad (5)$$

Where N' is the number of representative frames.

The computation of representative frames is based on a method proposed earlier in [6] with some modifications to reduce the processing time. The entire video sequence is first partitioned in to a few large segments of fixed length. The clustering method proposed in [6] is then applied to each of these large segments. The underlying assumption is that these partitions can be treated independently without degrading the overall performance of clustering for representative frame extraction, and our experiments indicate that this indeed is reasonable.

For each segment, the video is further divided into units each consisting of three consecutive frames. For each unit, the corresponding unit change is computed using (3). The units are then sorted based on their unit change values. During each iteration of the clustering process, units that have small unit changes as given by (3) are merged to form larger units. The algorithm iteratively converges to the desired number of units (5%, 10%, etc., depending on the level) at each level in an adaptive way based on the motion intensity change within units. The representative frames—the first and last frames of each unit—at each level are stored for indexing purpose. Details of the algorithm can be found in [6].

3. EXTRACTION OF MOTION ACTIVITY DESCRIPTORS IN DIFFERENT LEVELS

After the segmentation process, we obtain a representation of video in hierarchical levels, from coarse to fine, each consisting of a different number of video units. These units are then characterized using the following three motion activity descriptors: motion intensity of a P-frame, motion intensity histogram of a video unit, and spatial motion activity.

3.1. Motion intensity

We quantize motion intensity into levels for coding purposes. We use vector quantization methods to transform $G_u(\alpha)$ into N_l quantized intensity levels. The codebook of N_l entries is extracted from the $G_u(\alpha)$ data set first, then $G_u(\alpha)$ values are indexed using the code book. In our experiments, we set N_l =5. Therefore the motion intensity of a scene can be characterized by an intensity level $L = i$, where $i = 1,2,3,4,5.$

3.2. Motion intensity histogram

Motion intensity histogram (MIH) is used to represent the temporal statistics of motion intensity. Given a video unit, the corresponding motion intensity histogram of the unit is defined as: MIH=$[p_0, p_1, p_2, p_3, \ldots p_{N_l}]$, where p_i is the percentage of the quantized motion intensity corresponding to the *i*-th quantization level, and $\sum_{i=1}^{N_l} p_i = 1$. Here we set N_l =5. Within a short video segment, the MIH captures the human perception of motion quite well, as demonstrated by our experimental results.

3.3. Spatial motion activity descriptor

We use the magnitude of MPEG encoded motion vectors to form the spatial descriptor [2]. The extraction is as follows. For a given P frame, the average of the motion vector magnitude per macro-block of the frame is used to threshold the macroblocks into zero and non-zero-type based on their magnitudes. Then we compute lengths of runs of zeroes in the frame using a raster-scan order. Next, we classify the run-lengths into three categories, short, medium and long, that are normalized with respect to the object/frame width. In this case we have defined the short runs to be 1/3 of the frame width or lower, the medium runs to be greater than 1/3 the frame width and less than 2/3 of the frame width, and the long runs to be all runs that are greater than or equal to the frame width. N_{sr} is the number of short runs, and N_{mr} , N_{lr} , are similarly defined. Generally, the spatial descriptor can be applied to all the P frames in a video unit.

4. APPLICATION OF MOTION ACTIVITY FOR VIDEO INDEXING AND FILTERING

It is intuitive that different attributes are suitable at different granularities of the video. For instance, the spatial attribute is best used to describe a spatially homogenous unit of video and is not meaningful when applied to, say, an hour of video. We describe the utility of each attribute at each level in Table 1.

4.1. Application to video indexing

Descriptor/Level	Level1 (1%)	Level2 (5%)	Level3 (10%)	Level4 (20%)
Motion Intensity Average	Poor	Poor	Poor	Good
Motion Intensity Histogram	Good	Good	Good	Poor
Spatial descriptor	Poor	Good	Good	Good

Table 1. Expected performance of three descriptors at different levels.

Units	Precision	
	MIH	Spatial
News	63%	60%
Sports	80%	75%
Drama	30%	20%

Table 2. Subjective test results.

To test the descriptor, ninety-two units from 18 videos of news, sports, and drama from MPEG-7 data set are extracted. They are chosen to be 5% of their original video lengths, and are categorized into five groups based on MIH. Six subjects take the test to categorize them as well. In the case of spatial descriptor, the first frame of each unit is used for spatial descriptor computation. Table 2 lists the subjective test results. The results indicate that descriptors such as the MIH and spatial descriptors are best used with sports and news content. This is consistent with the fact that the semantics and motion features are strongly correlated in sports, moderately correlated in news, and not well correlated in drama.

An example of searching for similar video units based on the MIH descriptor is shown in Figure 2. The number of units is five percent of video. The first P frame from each unit is used to represent the whole unit. The query frame in Figure 2(a) shows a scene where a football game starts and there is very little motion in the scene. The five retrieved units, ranked in order from 1 to 5, are displayed in the figure. The first number in each rank bracket indicates the filtering result using the MIH descriptor. In general, we observe that scenes with similar motion patterns are retrieved from the video.

Figure 3 shows the experimental results based on the spatial motion activity descriptor. The first P frame from each video clip is shown in the figure. The query is an anchorperson in the news. We observe that the scenes with the anchorperson showing similar gestures are retrieved from the video stream.

4.2. Application for video filtering

The motion activity descriptor provides a unique way to segment content into semantically distinct units, and thus enables the user to get to the desired content by filtering or skipping over undesired content at different levels.

For example, sitcoms can be easily distinguished from high-motion-content sports scenes such as a soccer game. The temporal histogram of an hour of soccer video would have a high percentage of high action, which would help rule it out as typical drama content. Similarly the motion intensity descriptor helps us to directly reach the high action parts of a sports video or to skip over them as needed. However, the motion activity descriptor is not so useful for intra-program browsing, such as

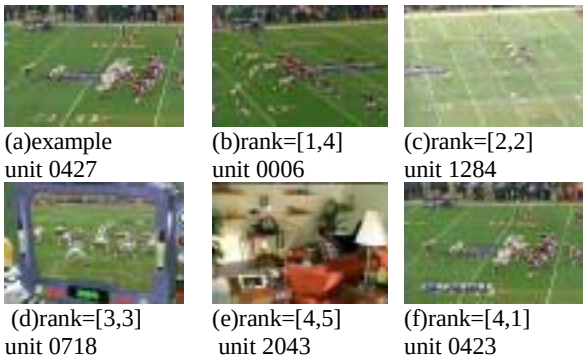

Figure 2. Motion intensity Histogram for indexing and filtering.

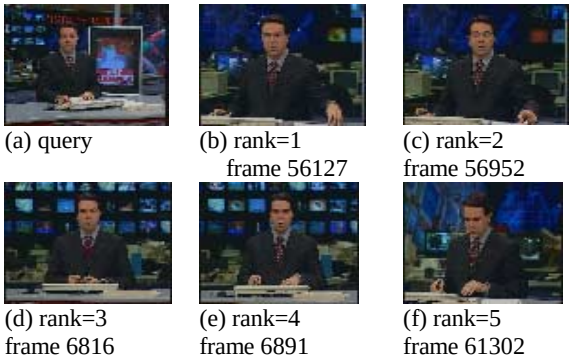

Figure 3. Spatial motion activity descriptor for indexing and searching.

searching within a sitcom program, since the gross motion characteristics of drama content do not change much.

For a sports or news video, MIH can serve to filter at the coarser levels of video. Once we have located the program of interest, the spatial attribute is useful in effectively locating similar activities among the video segments. Figure 2 gives results to show how to use such filtering technique. As discussed in the last section, MIH helps to filter out the five candidate units. Then they are used for further spatial filtering. The spatial descriptor for all P frames in all candidate units and the query unit are computed. The distance between a candidate and the query is computed as the smallest distance between two spatial descriptors, one from the query P frames and the other from the candidate P frames. Then we can reorder the candidates based on their distances to the query. The second number in each rank bracket indicates the result after the spatial filtering. Unit_0423 is our target, but it is the last one among the MIH filtering results. After further spatial processing, it moves to rank 1 as expected.

5. CONCLUSIONS

A novel method is proposed for motion activity extraction in hierarchical-levels for video indexing and filtering. To extract motion activity information, an MPEG (MPEG-1/2) video is first adaptively segmented into hierarchical levels based on P-frame motion information. The motion intensity, motion intensity histogram, and spatial motion descriptors are then computed to represent different levels of video. The descriptors from different levels are used selectively in different applications. The descriptor can be efficiently computed since only P frames of a video are processed and the descriptor can be processed directly in the MPEG compressed domain. Even though the MPEG-1/2 formats are discussed here, the method can be applied to other formats such as H.263 that have forward motion compensation..

This motion activity descriptor is useful in applications such as video indexing and filtering. While it is true that that the semantics and motion features are significantly correlated in sports and news video, the motion activity descriptor is still a low level descriptor. We are currently exploring the use of this descriptor together with other features such as color and texture for effective semantic level video indexing.

6. ACKNOWLEDGEMENT

This research is in part supported by the following grants/awards: Office of Naval Research grant #N00014-01-1-0391, an NSF instrumentation award #EIA-9986057, and an NSF infrastructure award #EIA-0080134.

7. REFERENCES

[1]. S. –F. Chang, W. Chen, H. Meng, H. Sundaram and D. Zhong, "A Fully Automated Content-Based Video Search Engine Supporting Spatiotemoral Queries," *IEEE Trans. on Circuits and Systems for Video Technology*, 8(5), pp.602-615, 1998.

[2]. A. Divakaran, H. Ito, H. Sun, P. Akella, P. Bouklee, A. Vetro, and T. Poon, "A Bit Allocation Based Descriptor for MPEG-4/2/1 Compressed Video Sequences," *ISO/IEC/SC29/WG11 MPEG99/P002*, 1999.

[3]. Y.Deng and B.S.Manjunath, "NeTra-V: toward an object-based video representation," *IEEE Transactions on Circuits and Systems for Video Technology*, .8(5), p.616-27, 1998.

[4]. B. G. Haskell, A. Puri and A. N. Netravali, "Digital Video: An Introduction to MPEG 2," Chapman and Hall, 1997.

[5]. H.R.Naphide, T.S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE Transactions on Multimedia*, 3(1), p.141-51, 2001.

[6]. X. Sun, M. Kankanhalli, Y. Zhu and J. Wu, "Content-Based Representative Frame Extraction for Digital Video," *Proc. ICMCS*, pp. 190-194, 1998.

[7]. X. Sun, B. S. Manjunath, P. Wu, Y. Deng, "Motion Quantized Alpha Histogram as a Video Unit Descriptor," *ISO/IEC JTC1/SC29/WG11/P75*, 1999.

[8]. X. Sun, J. Foote, D. Kimber, B. S. Manjunath, "Panoramic Video Capturing and Compressed Domain Virtual Camera Control," *Proc. ACM Multimedia*, pp. 329-338, 2001.

[9]. H. Zhang, C. Y. Low, and S. W. Smoliar, "Video parsing and browsing using compressed data," *Multimedia Tools and Applications*, 1(1): pp.89-111, 1995.

[10].http://www.almaden.ibm.com/cs/cuevideo

[11].http://www.virage.com/