MOTION ACTIVITY-BASED EXTRACTION OF KEY-FRAMES FROM VIDEO SHOTS

Ajay Divakaran, Regunathan Radhakrishnan and Kadir A. Peker

Mitsubishi Electric Research Laboratories
571 Central Avenue, Murray Hill, NJ 07974, USA
(ajayd,regu,peker@merl.com)

ABSTRACT

We describe a key-frame extraction technique based on the intuition that the higher the motion the more the key-frames required for summarization. We experimentally verify that the intensity of motion activity directly indicates the summarizability of the video segment, by using the MPEG-7 [1] motion activity descriptor and the fidelity measure described in [2]. We obtain the key-frames by dividing the shot in parts of equal cumulative motion activity, and then selecting the frame located at the half-way point of each sub-segment. Furthermore, we establish an empirical relationship between the motion activity of a segment and the required number of key-frames. We thus provide a unique and rapid way to find out the required number of key-frames and compute them. Our scheme is much faster than conventional color-based key-frame extraction schemes since it relies on simple computation and compressed domain extraction. It is close to the theoretical optimum in accuracy.

1. INTRODUCTION

Past work on finding key-frames of a shot (see [3] for a detailed review) has mostly focussed on using color features. There has been some work on video abstraction based on motion features such as Wolf [4]. In this paper we establish a novel concept viz. that the intensity of motion activity of a video segment is a measure of its "summarizability." We do so by studying the variation of the fidelity of a single key-frame with change in the intensity of motion activity as defined by the MPEG-7 video standard [1]. We then propose a novel technique for extraction of key-frames from shots based on the aforementioned measure of summarizability. We also propose a version of our technique that provides key-frames progressively. Our technique provides summarization that is numerically and visually comparable with the best existing techniques and relies on computationally simple motion feature extraction in the compressed domain, and is thus much simpler than other techniques.

2. MOTIVATION

2.1 Key Frame Extraction from Shots

An initial approach was to choose the first frame of a shot as the key-frame. It is a reasonable approach and works well for low-motion shots. However, as the motion becomes higher, the first frame is increasingly unacceptable as a key-frame. Many other subsequent approaches have built upon the first frame by using additional frames that significantly depart from the first frame in addition to the first frame. Other approaches rely on clustering and other computationally intensive analysis. All of the approaches mentioned up to now do not make use of motion features and are computationally intensive. The reason for using color is that it enables a reliable measure of change from frame to frame. However, motion-compensated video also relies on measurement of change from frame to frame, which motivates us to investigate schemes that use motion vectors to sense the change from frame to frame in a video sequence. Furthermore, motion vectors are readily available in the compressed domain hence offering a computationally attractive avenue. Our approach is similar to Wolf's approach in that we also make use of a simple motion metric and in that we do not make use of fixed thresholds to decide which frames will be key-frames. However, unlike Wolf, instead of following the variation of the measure from frame to frame, we propose that the simple **shot-wide** motion metric, the MPEG-7 intensity of motion activity descriptor, is a measure of the summarizability of the video sequence.

2.2 The Fidelity of a Set of Key Frames

The fidelity measure [2] is defined as the Semi-Hausdorff distance between the set of key-frames S and the set of frames R in the video sequences. A practical definition of the Semi-Hausdorff distance is as follows:

Let the key frame set consist of m frames S_i i=1..m, and let the set of frames R contain n frames R_i i=1..n. Let the distance between two frames S_i and R_i be $d(S_i,R_i)$. Define d_i for each frame R_i as

$$d_i = \min(d(S_k, R_i)), k = 0..m$$

Then the Semi-Hausdorff distance between S and R is given by

$$d_{sh}(S,R) = \max(d_i), i = 1..n$$

Most existing dissimilarity measures satisfy the properties required for the distance over a metric space used in the above definition. In this paper, we use the color histogram intersection metric proposed by Swain and Ballard (See [2]).

2.3 The MPEG-7 Motion Activity Descriptor

The MPEG-7 [1] motion activity descriptor attempts to capture human perception of the "intensity of action" or the "pace" of a video segment. For instance, a goal scoring moment in a soccer game would be perceived as a "high action" sequence by most if not all human viewers. On the other hand, a "head and shoulders" sequence of a talking person would certainly be considered a "low action" sequence by most. The MPEG-7 motion activity descriptor has been found to accurately capture the entire range of intensity of action in natural video. It uses quantized standard deviation of motion vectors to classify video segments into five classes ranging from very low to very high intensity.

2.4 Motion Activity as a Measure of Summarizability

In this paper, we hypothesize that since high or low action is in fact a measure of how much a video scene is changing, it is a measure of the "summarizability" of the video scene. For instance, a high speed car chase will certainly have many more "changes" in it compared to say a news-anchor shot, and thus the high speed car chase will require more resources for a visual summary than would a news-anchor shot.

Unfortunately, there are no simple objective measures to test such a hypothesis. However, since change in a scene often also involves change in the color characteristics as well, we first try to investigate the relationship between color-based fidelity as defined in 2.2, and intensity of motion activity. Let the key frame set for shot A be S_A and that for shot B be S_B. If S_A and S_B both contain the same number of key frames, then our hypothesis is that if the intensity of motion activity of shot A is greater than the intensity of motion activity of shot B, then the fidelity of S_A is less than the fidelity of S_B.

3. EXPERIMENTAL PROCEDURE AND RESULTS

3.1 Establishing the Hypothesis

We extract the color and motion features of news video programs from the MPEG-7 test-set, which is in the MPEG-1 format. We first segment the programs into shots. For each shot, we then extract the motion activity features from all the P-frames by computing the standard deviation of motion vector magnitudes of the macro-blocks of each P frame, and a 64 bin RGB Histogram from all the I-frames, both in the compressed domain. Note that intra-coded blocks are considered to have zero motion vector magnitude. We then compute the motion activity descriptor for each I-Frame by averaging those of the previous P-frames in the Group of Pictures (GOP). The I-Frames thus all have a histogram and a motion activity value associated with them. The motion activity of the entire shot is got by averaging the individual motion activity values computed above. From now on, we treat the set of I-frames in the shot as the set of frames R as defined earlier.

The simplest strategy for generating a single key frame for a shot is to use the first frame, as mentioned earlier. We thus use the first I-frame as the key frame and compute its fidelity as described in 2.2. We find empirically that a key frame with Semi-Hausdorff distance at most 0.2 is of satisfactory quality, by analyzing examples of "talking head" sequences. We can therefore classify the shots into two categories, those with key frames with d_{sh} less than or equal to 0.2 i.e. of acceptable fidelity and those with key frames with d_{sh} greater than 0.2, i.e. unacceptable fidelity. Using the MPEG-7 motion activity descriptor, we can also classify the shots into five categories ranging from very low to very high activity. We then find the percentage duration of shots with d_{sh} greater than 0.2 in each of these categories for the news program news1 (Spanish News) and plot the results in Figure 1. We can see that as the motion activity goes up from very low to very high, the value of d_{sh} also increases consistently. In other words, the summarizability of the shots goes down as their motion activity goes up. Furthermore, the fidelity of the single key frame is acceptable for 90% of the shots in the very low intensity of motion activity category. We find the same pattern with other news programs. We thus find experimental evidence that with news program content, our hypothesis is valid. Since news programs are diverse in content, we would expect this result to apply to a wide variety of content. Since we use the MPEG-7 thresholds for motion activity, our result is not content dependent.

3.2 A Motion Activity based Non-Uniform Sampling Approach to Key Frame Extraction

If as per section 3.1 intensity of motion activity is indeed a measure of the change from frame to frame, then over time, the cumulative intensity of motion activity must be a good indication of the cumulative change in the content.

Recall that in our review of previous work we stated that being forced to pick the first frame as a key-frame is disadvantageous. If the first frame is not the best choice for the best first key-frame, schemes that use it as the first key-frame such as [5] start off at a disadvantage. This motivates us to find a better single key-frame based on motion activity. If each frame represents an increment in information then the last frame is at the maximum distance from the first. That would imply that the frame at which the cumulative motion activity is half the maximum value is the best choice for the first key-frame. We test this hypothesis by using the frame at which the cumulative motion activity is half its value for the entire shot as the single key-frame instead of the first key frame for the Spanish News sequence and repeating the experiment in the previous section. We find that that the new key-frame choice outperforms the first frame, as illustrated in Figure 1. Since previous schemes have also improved upon using the first frame as a key-frame, we need to compare our single key-frame extraction strategy with them. For each shot, we compute the optimal single key-frame as per the fidelity criterion mentioned in section 2.2. We compute it by finding the fidelity of each of the frames of the video, and then finding the frame with the best fidelity. We use the fidelity of the aforementioned optimal key-frame as a benchmark for our key-frame extraction strategy by measuring the difference in d_{sh} between the optimal key-frame obtained through the exhaustive computation mentioned earlier and the key-frame obtained through our proposed motion-activity based strategy. We carry out a similar comparison for the first-frame based strategy as well. We illustrate our results in Table 1. Note that our strategy produces key-frames that are nearly optimal in fidelity. Furthermore, the quality of the approximation degrades as the intensity of motion activity increases. In other words, we find that our strategy closely approximates the optimal key-frame extraction in terms of fidelity while using much less computation.

This motivates us to propose a new nearly optimal strategy, which is very similar to the activity-based sampling proposed in [6] as follows. To get n key-frames, divide the video sequence into n equal parts on the cumulative motion activity scale. Then use the frame at the middle of the cumulative motion activity scale of each of the segments as a key-frame, thus getting n key-frames. Note that our n key-frame extraction strategy scales linearly with n unlike the exhaustive computation described earlier, which grows exponentially in complexity because of the growth in the number of candidate key-frame combinations. It is for this reason that we do not compare our n-frame strategy with the exhaustive benchmark.

A Simple Progressive Modification

Since our key-frame extraction is not progressive, we propose a progressive modification of our technique. We start with the first frame, and then choose the last frame as the next key-frame because it is at the greatest distance from the first frame. We carry this logic forward as we compute further key-frames by choosing the middle key-frame as the third key-frame, and so on recursively. The modified version is slightly inferior to our original technique but has the advantage of being progressive. In figure 2, we illustrate a typical result. We have tried our approach with several news programs from different sources.

4. DISCUSSION AND CONCLUSION

We hypothesized that the intensity of motion activity of a video segment is an indication of its summarizability. Our experimental results with diverse content show that our hypothesis is correct. Our results motivated a novel key frame extraction strategy that relies on activity based non-uniform sampling of frames. It is computationally extremely simple in the compressed domain. It gives visually acceptable results that are nearly optimal over a wide variety of news programs and content. Its progressive version is a natural generalization of picking the first frame as the key frame, and can also help speed up color-based summarization. In future work, we plan to investigate further application and extension of our summarization scheme to consumer and surveillance video. Such application should certainly involve inclusion of other cues such as other visual features, audio, text from closed captions, meta-data, and embedded captions etc. We have built a video browsing system based on our approach.

5. REFERENCES

1. S Jeannin and A. Divakaran, MPEG-7 visual motion descriptors, IEEE Transactions on Circuits and Systems for Video Technology, Vol 11, No. 6, pp. 720-724, June 2001.
2. H.S. Chang, S. Sull and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp. 1269-1279, December 1999.
3. A Hanjalic and H. Zhang, An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis, IEEE Trans. On Circuits and Systems for Video Technology, Vol. 9, No. 8, December 1999.
4. W. Wolf, Key frame selection by motion analysis, ICASSP '96, Atlanta, GA, pp. 1228-1231, 1996.
5. M.M. Yeung and B. Liu, "Efficient Matching and Clustering of Video Shots," Proc. IEEE ICIP, pp. I.338-I.341, Washington D.C., 1995.
6. K. Peker, A. Divakaran and H. Sun, Constant pace skimming and temporal sub-sampling of video using motion activity, Proc. IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece, October 2001.

7. A. Divakaran, R. Radhakrishnan and K. Peker "Video Summarization with Motion Descriptors," Journal of Electronic Imaging, October 2001.

MPEG-7 Intensity of Motion Activity	Average Difference between d_{sh} of first frame and optimal key-frame	Average Difference between d_{sh} of proposed key-frame and optimal key-frame	Number of Shots
Very Low	0.0245	0.0143	52
Low	0.0442	0.0281	134
Medium	0.0877	0.0518	48
High	0.1064	0.0126	2
Very High	-----	------	0
Overall average difference in d_{sh} over entire program	0.0492	0.0297	

Table 1: Comparison of proposed key-frame extraction strategy with exhaustively computed optimal key-frame for news1, Spanish News from MPEG-7 test-set.

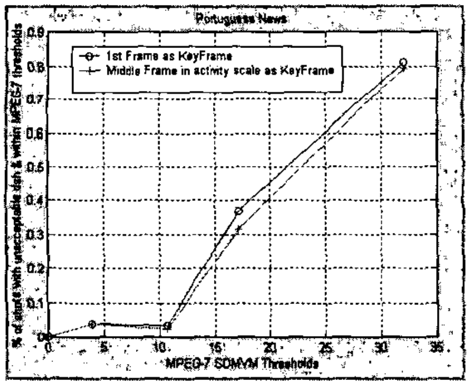

Figure 1: Verification of Hypothesis and choice of single key-frame Motion activity (Standard Deviation of Motion Vector Magnitude) Vs percentage duration of unacceptable Shots (Spanish News from MPEG-7 Test Set news1.mpg)

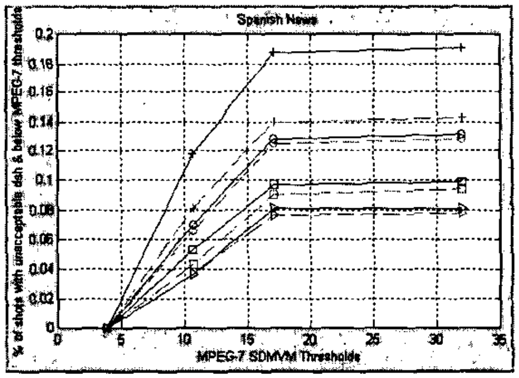

Figure 2: Motion activity (Standard Deviation of Motion Vector Magnitude) Vs % dur.of unacc. Shots (Spanish News from MPEG-7 Test Set) The firm line represents the "optimal" key-frame strategy while the dotted line represents the progressive key-frame extraction strategy. Each shape represents a certain number of key-frames, the + represents a single frame, the circle two frames, the square three frames and the triangle five frames.