MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Adaptive Fast Playback-Based Video Skimming Using a Compressed-Domain Visual Complexity Measure

Kadir A. Peker, Ajay Divakaran

TR2004-060 December 2004

Abstract

We present a novel compressed domain measure of spatio-temporal activity or visual complexity of a video segment. The visual complexity measure indicates how fast a video segment can be played within human perceptual limits. We present an adaptive smart fast-forward based video skimming method where the playback speed is varied based on the visual complexity. Alternatively, spatio-temporal smoothing is used to reduce visual complexity for an acceptable playback at a given playback speed. The complexity measure and the skimming method are based on early vision principles, thus they are applicable across a wide range of content type and applications. It is best suited for low temporal compression instant skims. It preserves the temporal continuity and eliminates the risk of missing an important event. It can be extended to include semantic inputs such as face or event detection, or can be a presentation end to semantic summarization.

ICME 2004

Adaptive Fast Playback-Based Video Skimming Using A Compressed-Domain Visual Complexity Measure

Kadir A. Peker and Ajay Divakaran
Mitsubishi Electric Research Laboratories
201 Broadway, Cambridge, MA 02139, USA
+1 - 617 621 7500
{peker, ajayd}@merl.com

Abstract

We present a novel compressed domain measure of spatio-temporal activity or visual complexity of a video segment. The visual complexity measure indicates how fast a video segment can be played within human perceptual limits. We present an adaptive "smart fast-forward" based video skimming method where the playback speed is varied based on the visual complexity. Alternatively, spatio-temporal smoothing is used to reduce visual complexity for an acceptable playback at a given playback speed. The complexity measure and the skimming method are based on early vision principles, thus they are applicable across a wide range of content type and applications. It is best suited for low temporal compression instant skims. It preserves the temporal continuity and eliminates the risk of missing an important event. It can be extended to include semantic inputs such as face or event detection, or can be a presentation end to semantic summarization.

1. Introduction

Past approaches to summarization include clustering video frames and selecting representative frames from clusters, using a measure of change in the video content along time and selecting representative frames whenever the change is significant, and approaches based on assigning some significance measure to the parts of the video and subsequently filtering less significant parts [4]. Video summaries have been presented as key-frames, video skims, or mosaics. See [3] for a thorough review.

We have previously presented an adaptive fast playback-based video summarization framework [5][6] wherein the playback rate was modified so as to maintain a constant "pace" throughout the content. We assumed that the motion activity descriptor, which is the average magnitude of the motion vectors in mpeg video, provides a measure of visual information rate or "pace". Then we maintain a constant "pace" by modifying the playback rate, hence linearly changing the motion activity. The process is similar to a bandwidth allocation scheme, where information rate is given by motion activity and the human visual system has a certain channel bandwidth. Note that no use is made of spatial features such as texture.

In this paper, we introduce a novel compressed domain "visual complexity" feature that is a function of the spatial complexity (texture) as well as the temporal complexity (motion activity) of video. We then describe changing the playback rate and applying spatio-temporal smoothing adaptively, based on visual complexity, for an effective fast playback.

Our approach is based on early vision principles and thus does not claim semantic significance. It thus has applicability across various content types and application contexts. The time compression ratio is low to moderate, suitable for instant video skims at any point in the video in the form of a "smart" fast forwarding or rewinding. In this type of video skimming, the temporal continuity is preserved; the risk of losing important segments is also eliminated.

2. Human Visual System and the Visual Complexity

The highest speed at which you can playback a video segment with acceptable comprehension of its content is a function of a number of factors, including the scene complexity, the semantic elements in the scene, the familiarity of those elements, the processing capacity of the visual system, etc. However, modeling the semantic and the memory aspects of human vision

system is very difficult. Instead, we develop a visual complexity measure based on early vision models.

We postulate that the visual complexity of a video segment is proportional to its spatio-temporal bandwidth. The human visual system is sensitive to stimuli only in a certain spatio-temporal window, called "the window of visibility" [1]. That is, we cannot see beyond a certain spatial resolution or temporal frequency limit.

The highest spatial resolution we can see ranges from 6 to 60 cycles/degree depending on viewing conditions [1][2], which is higher than what most current display systems provide (e.g. HDTV 30 cycles/degree). Thus, the bottleneck in visual complexity is usually not the spatial bandwidth. The temporal frequency limit reported under the same conditions is around 30 Hz, which is comparable to TV (25 or 30) and film (24) frame rates. In a fast playback of digital (i.e. time sampled) video, the temporal bandwidth of the video is stretched wider, where some of the action falls beyond the window of visibility. If speed up is achieved by dropping frames, aliasing occurs as well. We show that the temporal bandwidth of the speeded video, which determines the perceived quality of the fast playback, is a function of the spatial bandwidth ("texturedness") and the motion activity of the original video.

3. Estimation of Temporal Bandwidth

Let us assume a one-dimensional sinusoidal signal with frequency w_s moving in the positive ***x*** direction with speed v. The amplitude variation in time at a fixed ***x***-position is also sinusoidal, with temporal frequency $w_t = v \cdot w_s$. In general, a 1-D signal translating in time has a spatio-temporal spectrum lying on a line passing through the origin. A translating band-limited signal with a spatial bandwidth *U*, has a spatio-temporal transform extending on a line from $(U, -v \cdot U)$ to $(-U, v \cdot U)$.

In the 2-D case, we show that the temporal frequency of a moving sinusoid is given by the dot product of the frequency vector and the velocity vector. Figure 1 shows the 2-D sinusoid $\cos(2\pi \frac{1}{N} x + 2\pi \frac{4}{N} y)$. The frequency along the x-axis is $f_x = 0.5$, and the frequency along the y-axis is $f_y = 2$. We represent this sinusoid with a frequency vector $\vec{\mathbf{f}}$ = (0.5, 2). If this 2-D sinusoid is translating with motion $\vec{\mathbf{v}} = (v_x, v_y)$, then the temporal frequency (i.e. the frequency of change at a fixed location in time) is,

$$w_t = \vec{\mathbf{f}} \cdot \vec{\mathbf{v}} .$$

Note that the temporal frequency is highest when the motion is perpendicular to the wave front, and is zero when it is parallel to it.

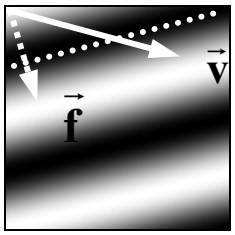

Figure 1. A 2-D sinusoid with a frequency vector **f,** and a motion vector **v** showing its translation velocity.

4. Computation in the MPEG-1/2 Compressed Domain

We will find the temporal bandwidth estimate (i.e. the visual complexity) for each DCT block in compressed video. Each DCT block is a superposition of several 2-D sinusoids moving with block motion vector $\vec{\mathbf{v}} = (v_x, v_y)$. The basis functions of the DCT transformation are in the form;

$$\cos(\frac{\pi k_x (2x+1)}{2N}) \cdot \cos(\frac{\pi k_y (2y+1)}{2N})$$

$$= \cos(2\pi \frac{k_x}{2N} x + 2\pi \frac{k}{4N}) \cdot \cos(2\pi \frac{k_y}{2N} y + 2\pi \frac{k}{4N}),$$

which is the multiplication of two 1-D sinusoids with frequencies $\frac{k_x}{2}$ and $\frac{k_y}{2}$, whereas a 2-D sinusoidal grating with a frequency $\vec{\mathbf{f}} = (f_x, f_y)$ is given as;

$$\cos(2\pi \frac{f_x}{N} x + 2\pi \frac{f_y}{N} y) \cdot$$

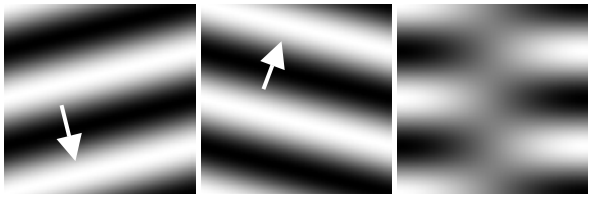

Figure 2. The two 2-D sinusoids that make up a DCT basis when summed up.

By the cosine expansion of the DCT basis as;

$$= \frac{1}{2}\left[\begin{array}{l}\cos\left(2\pi\frac{k_x}{2N}x + 2\pi\frac{k_y}{2N}y + 2\pi\frac{k_x+k_y}{4N}\right) \\ +\cos\left(2\pi\frac{k_x}{2N}x - 2\pi\frac{k_y}{2N}y + 2\pi\frac{k_x-k_y}{4N}\right)\end{array}\right]$$

each DCT basis is a superposition of two 2-D sinusoids, one with spatial frequency $\vec{\mathbf{f}}_1 = (\frac{k_x}{2}, \frac{k_y}{2})$ and the other with $\vec{\mathbf{f}}_2 = (\frac{k_x}{2}, -\frac{k_y}{2})$ (Figure 2). As each of these components move with the block motion vector $\vec{\mathbf{v}} = (v_x, v_y)$, we observe two temporal frequency components for each DCT coefficient. The corresponding temporal frequencies are given by;

$$\omega_1 = \vec{\mathbf{f}}_1 \cdot \vec{\mathbf{v}}_1 = \frac{k_x}{2}v_x + \frac{k_y}{2}v_y, \text{ and}$$

$$\omega_2 = \vec{\mathbf{f}}_2 \cdot \vec{\mathbf{v}}_2 = \frac{k_x}{2}v_x - \frac{k_y}{2}v_y \text{ (cycles/block)}$$

Converting to cycles-per-pixel and using the absolute values of the temporal frequencies, the final form of the temporal frequency components contributed by each DCT coefficient is;

$$\omega_1 = \frac{|k_x v_x + k_y v_y|}{16}, \text{ and } \omega_2 = \frac{|k_x v_x - k_y v_y|}{16} \text{ cycle/frame;}$$

The magnitude of the temporal frequency components ω_1 and ω_2 contributed by each DCT coefficient is equal to half of the energy of that DCT coefficient due to the DCT cosine expansion.

4.1. Motion Vector and DCT Estimation

We discard low-texture blocks since the motion vectors are less reliable for those blocks [7]. Note that low-texture blocks are expected to have low visual complexity, hence the effect on visual complexity computation is minimal. We then apply median filtering to further eliminate spurious motion vectors.

We can compute the DCT coefficients of P frame blocks by applying motion compensation or estimate without decoding. Alternatively, we can consider the motion vectors from an I-frame to the following P frame as the motion of blocks on a non-regular grid in the I-frame. Then we can interpolate the motion vector field for the regular DCT block grid. We find the latter approach faster and easier to implement.

4.2. Spatio-temporal Complexity of a Video Segment

We first create a histogram of the temporal frequencies and their energy, computed as described in the previous section. Each DCT block has a histogram, and the frame has a combined histogram, which is an approximate temporal spectrum of the video segment around that frame. We define the visual complexity as a number that captures the effective temporal bandwidth of the video segment, computed as a weighted mean or a percentile from the temporal frequency histogram. We compute a complexity for each DCT block (see Figure 3 for a sample frame) and average (possibly weighted by a region significance factor) for the whole frame. Note that we defined the visual complexity as the spatio-temporal bandwidth of the video segment, however, the spatial bandwidth is constant and is always within the window of visibility as described earlier. Also note that, the temporal bandwidth can be computed through a 3-D FFT. However, this is impractical due to the complexity and the buffer requirements. Our method provides a practical and feasible compressed domain approximation based on the piece-wise linear motion assumption.

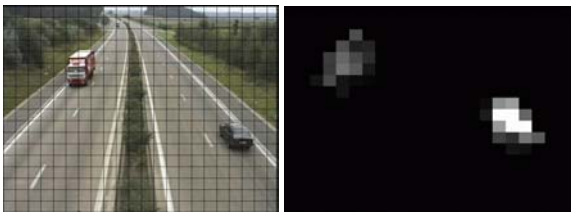

Figure 3. A frame from an MPEG-7 test video, and its visual complexity for each DCT block.

5. Adaptive Fast Playback

Adaptive playback can be formulated as a time warping of a video stream $V_0(x,y,t)$ into $V_1(x,y,t)$:

$$V_1(x, y, w(t)) = V_0(x, y, t)$$

where $w(t): [0, T_0] \rightarrow [0, T_1]$ is a non-linear mapping of the interval $[0, T_0]$ onto $[0, T_1]$, and T_0 and T_1 are the duration of V_0 and V_1, respectively. For digital video, the warping can be implemented either by adaptively changing the playback rate (play frame *i* at time $w(t_i)$ instead of t_i) or by adaptively dropping frames preserving the frame rate. The latter is better suited for practical applications. We describe a simple implementation through accumulation and thresholding of the motion activity in [6].

When the video playback is speeded up, the motion vectors are scaled up proportionately; hence the visual complexity increases. In order to utilize the assumed

visual bandwidth most efficiently, we speed up the video frames inversely proportional to their visual complexity, thus maintaining a constant visual complexity throughout the video. We are also working on a number of smoothness constraints that will make the playback more pleasant, such as discrete levels of speed, smooth transitions between levels, minimum required time at a given speed, maximum speed change at a time, etc.

The visual complexity is a function of the motion vectors (controlled through the playback speed), and the spatial frequencies present, or the level of fine texture in the video. The latter can be controlled by applying spatio-temporal filtering (e.g. motion blur) to reduce the visual complexity so that a video segment can be played faster.

6. Discussion and Conclusions

We presented an intuitive measure of visual complexity of a video segment that combines the spatial complexity and the amount of motion in the scene. We described a framework for skimming through video segments using visual complexity and adaptive fast playback.

The spatio-temporal complexity is a superset of the motion activity feature (See Figure 4) we presented in [6]. It reduces to motion activity when the spatial complexity of the scenes is the same for all the frames in the video. The new visual complexity extends the motion activity by introducing the spatial scene complexity as a variable. However, in certain applications such as sports highlights detection, we are primarily interested in the motion itself, hence the spatio-temporal complexity may not provide an advantage over motion activity.

We are working on integrating visual complexity based fast playback with semantic summarization methods. We are also considering background - foreground separation to avoid slowing video for unnecessary detail in the background.

7. References

[1] A. Watson, A. Ahumada, J Farrell, "Window of Visibility: a psychophysical theory of fidelity in time-sampled visual motion displays," J. Opt. Soc. Am. A, Vol. 3, No. 3, pp. 300-307, Mar 86.

[2] NDT Resource Center, Visual Acuity of Human Eye (http://www.ndt-ed.org/EducationResources /CommunityCollege/PenetrantTest/Introduction/visualacuity.htm

[3] A Hanjalic and H. Zhang, An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis, IEEE Trans. CSVT, Vol. 9, No. 8, December 1999.

[4] Y-F. Ma, L. Lu, H-J. Zhang, and M. Li, "A User Attention Model for Video Summarization," ACM Multimedia 02, pp. 533 – 542, December 2002.

[5] Divakaran, A.; Peker, K.A.; Radharkishnan, R.; Xiong, Z.; Cabasson, R., "Video Summarization Using MPEG-7 Motion Activity and Audio Descriptors", in *Video Mining*, Rosenfeld, A.; Doermann, D.; DeMenthon, D., Kluwer Academic Pub, Oct. 2003.

[6] K. A. Peker, A. Divakaran and H. Sun, "Constant pace skimming and temporal sub-sampling of video using motion activity," Proc. ICIP 2001, Greece.

[7] M. Pilu, "Motion re-estimation from raw MPEG vectors with applications to image mosaicing", SPIE Electronic Imaging Conference, Jan 1998.

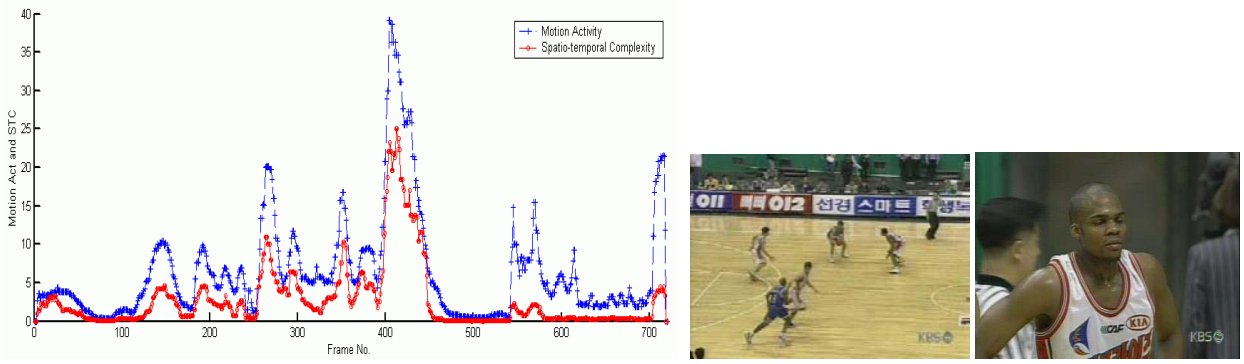

Figure 4. Motion activity and visual (spatio-temporal) complexity for a basketball video segment (MPEG7 testset). The two are similar except in the last part, which is a close up on a player. Visual complexity is lower here because the images are larger with larger spatial features compared to wide shots.