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

TR2003-33 June 2004

Abstract

Psychophysical experiments have shown that the human visual system is sensitive to visual stimuli only within a certain spatio-temporal window. the location of a moving image in the spatio-temporal space is determined by the spatial frequency content of image regions and their velocity. We present a novel compressed domain measure of spatio-temporal motion activity of a video segment that provides us with a criteria on how fast a video segment can be played within human perceptual limits. Alternatively, this measure allows us to determine the spatio-temporal filtering required for an acceptable playback of a video segment at a given fast playback speed. The spatio-temporal activity measure is computed in the compressed domain and allows for generation of instant skims through video content at any point forward using adaptive fast playback. The adaptive fast playback method using spatio-temporal complexity is based on early vision characteristics of the human visual system only, and thus independent of content type and semantics so it is applicable in a wide range of applications. It is best suited for low temporal compression summarization. A visual of the content is preserved at all times; hence the temporal continuity of the action is preserved, and the risk of missing an important event is eliminated as well. The user can switch between skim mode and regular playback at anytime or change the speed-up ratio of the fast playback. Our simulations on various types of video indicate that the presented video skimming and summarization method is effective and useful. Finally, the adaptive fast playback framework can be extended to include other inputs such as face detection, dialog detection, or semantic annotation. It can also be integrated with other summarization methods that try to capture the semantics.

IEEE International conference on Multimedia and Expo (ICME)

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
Psychophysical experiments have shown that the human visual system is sensitive to visual stimuli only within a certain spatio-temporal window. The location of a moving image in the spatio-temporal space is determined by the spatial frequency content of image regions and their velocity. We present a novel compressed domain measure of spatio-temporal motion activity of a video segment that provides us with a criteria on how fast a video segment can be played within human perceptual limits. Alternatively, this measure allows us to determine the spatio-temporal filtering required for an acceptable playback of a video segment at a given fast playback speed. The spatio-temporal activity measure is computed in the compressed domain and allows for generation of instant skims through video content at any point forward using adaptive fast playback. The adaptive fast playback method using spatio-temporal complexity is based on early vision characteristics of the human visual system only, and thus independent of content type and semantics, so it is applicable in a wide range of applications. It is best suited for low temporal compression summarization. A visual of the content is preserved at all times; hence the temporal continuity of the action is preserved, and the risk of missing an important event is eliminated as well. The user can switch between skim mode and regular playback at anytime or change the speed-up ratio of the fast playback. Our simulations on various types of video indicate that the presented video skimming and summarization method is effective and useful. Finally, the adaptive fast playback framework can be extended to include other inputs such as face detection, dialog detection, or semantic annotation. It can also be integrated with other summarization methods that try to capture the semantics.

Keywords
Video summarization, skimming, video content analysis, human visual system.

1. INTRODUCTION
In recent years, several video summarization approaches have been introduced. One of the approaches is based on reducing redundancy by clustering video frames and selecting representative frames from clusters [2][3][4]. Another approach is using a measure of change in the video content along time, and selecting representative frames whenever the change becomes significant [5][6]. Finally, there have been approaches based on assigning some significance measure to the parts of the video – usually based on criteria inspired from the human visual system – and subsequently filtering less significant parts [7].

In terms of the presentation style, we can identify two main categories of video summaries: still image-based summaries, and motion summaries. Many of the above approaches can be used to generate either of these types of summaries. Various other visualization options such as video mosaics have also been proposed.

We have previously presented an adaptive fast playback-based video summarization framework [10][9]. The playback rate was modified so as to maintain a constant "pace" through out the content. We assumed the motion activity descriptor, which is the average magnitude of the motion vectors in mpeg video, to provide a measure of the "pace" of the content. This approach can be viewed as a bandwidth allocation scheme, where we are given a measure of the "visual bandwidth" of the video and a channel bandwidth defined by the human visual system. Since the motion activity measure we use is linearly proportional with the playback rate of the video, we can linearly increase or decrease the visual bandwidth of the video by changing the playback rate. Hence, we achieve the optimum time-visual bandwidth allocation by adaptively changing the playback rate so as to have a constant motion activity during the playback of the video.

In this paper, we elaborate on the above approach first by providing a more refined investigation of, and a measure for the "visual bandwidth", or the visual complexity of a video segment. The visual complexity of a scene, which determines how fast the eye can follow the flow of action, is a function of spatial complexity as well as the temporal complexity. We introduce a novel compressed domain feature that combines these two factors, and base it on proposed models of the human visual system.

Equipped with a measure of visual complexity of a video scene, we present two alternative ways of using it in video skimming. One use is that, given a video content, we can determine the maximum rate at which each segment can be played back. The

other use is in determining how we need to filter the content to play it back at a given rate. This second approach makes use of the fact that the visual complexity is partly a function of the spatial complexity, which can be reduced by filtering out high frequency spatial components, e.g. spatio-temporal smoothing.

Note that the adaptive playback approach is essentially different from assigning a significance score to video segments. Although possible to extend that way, we do not use the visual complexity in deciding to exclude or include video segments, in a 1 or 0 fashion. For example, a low visual complexity does not mean that a certain segment is dispensable, but means that it requires less time to convey through the visual system.

Also note that the visual complexity measure does not imply any semantic inferences. The playback rate is adapted only to the low-level physical characteristics of the content and is based on the early vision stage of the human visual system, rather than the higher cognitive stages. In this aspect, adaptive fast playback-based video skimming is closer to a presentation method than a semantic content analysis approach. Hence, it is complimentary to any other summarization method. However the video to be presented is selected – either the full video or a summarized subset, it can be effectively viewed using adaptive fast playback.

We can illustrate the difference between the early vision approach of adaptive fast playback and other content based video summarization methods with a parallel example from image domain. For instance, a close-up on a key human face, or a text image with large letters and a white background can be semantically much more important than a high detail natural background. But the former two may require far fewer bits to represent than the last. Similarly, the adaptive playback method may assign more playback time to a semantically less important but visually complex scene, compared to a semantically more important but visually less demanding scene. Thus, our goal in this paper is determining the most efficient presentation for a piece of video through accelerated playback, regardless of its semantic content. We will discuss the integration with semantic summarization methods at the end of the paper.

The time compression ratio that can be achieved through adaptive fast playback is relatively smaller than the methods where you select a much smaller subset of the original content. This method is most suitable when the required time-compression ratio is not too high, and preserving the visual continuity is a desired feature. The instant availability of such a skimming feature during a regular playback is useful in seeking through a video segment and zeroing in on the parts that the user wants to see in detail. Preserving the continuity of the activity is also a desired feature in some applications. In certain cases, the user may prefer to flash through a video and see for him or herself what it is about – similar to thumbing through a book, rather than relying on an automatically selected set of disconnected pieces. Finally, fast playback can be used in cases where the reliability of automated methods is not sufficient or the risk associated with missing a desired segment is very high, e.g. in certain surveillance applications.

The visual complexity measure we present in this paper is an extension of the MPEG-7 motion activity feature so as to include the spatial domain as well. We use the terms visual complexity and spatio-temporal complexity interchangeably in the rest of the paper.

2. ADAPTIVE FAST PLAYBACK

We can view the adaptive changing of the playback frame rate of a video stream as time warping. We define $V_0(x,y,t)$ as the original video in continuous time, and $V_1(x,y,t)$ as the time-warped version of it with:

$$V_1(x,y,t) = V_0(x,y,w(t))$$
$$\text{where } w(t):[0,T_1] \rightarrow [0,T_0],$$
$$T_0 : \text{duration of } V_0,\ T_1 : \text{duration of } V_1,$$

is a warping function on the time axis. The original video is sampled at r frames per second, resulting in the discrete stream $V_0(x,y,f_f)$.

One way of implementing the adaptive playback rate is by changing the actual playback frame rate of the video. In this implementation, we compute the time instants that each frame of the original video V_0 should appear in the time warped video V_1, and display them at those times, resulting in a non-uniform time sampling of the time warped video $V_1(x,y,t)$. However, most practical video players do not support such a non-uniform playback rate. Another implementation problem is the high playback frame rates that would be required for speeded-up segments with this approach.

The second approach is to modify the effective playback rate by adaptively dropping frames, but playing the final video at a fixed frame rate. In this version, the presented video is uniformly sampled, but the original video (V_0) is non-uniformly sampled. A quantization-like effect causes the achieved frame rate and the desired frame rate to be different at times, unless interpolation of frames is used.

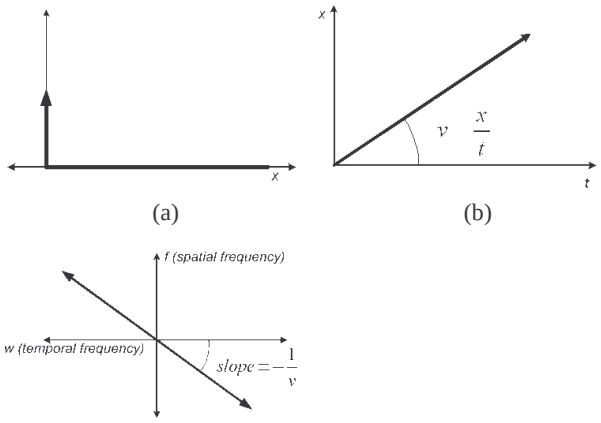
3. HUMAN VISUAL SYSTEM AND VISUAL COMPLEXITY

The speed at which you can playback a segment of video with acceptable comprehension of its content is a function of number of factors, possibly including the scene complexity, semantic elements in the scene, the familiarity of those elements and the scene, the processing capacity of the visual system, etc. One way of modeling the problem is through the definition of such a "frame processing time" of the human visual system. However, it is very difficult to model the semantic and the memory parameters in this function. An alternative approach is based on early vision models proposed in psychophysics area.

Research in psychophysics has shown that the human visual system is sensitive to stimuli only in a certain spatio-temporal window, called the window of visibility [1]. That is, we can not see beyond a certain spatial resolution or temporal frequency limit. Watson and Ahumada postulate that, for a time sampled video signal to be perceived the same as its continuous version, the two signals should look the same within the window of visibility, in the transform domain. This result gives us a trade-off relationship between the spatial bandwidth and the velocity of visual stimuli, i.e. 2-D images in the case of video, for them to preserve the same visual quality.

To illustrate these points, let us consider a 1-D signal in linear motion. Figure 1 shows an impulse signal moving left with speed

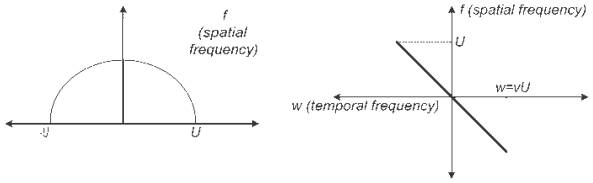
v, such that $x = v \cdot t$. This corresponds to a line in the x-t space. The Fourier transform of this signal is also a line, passing through the origin, with slope $-\frac{1}{v}$ [1]. In general, a 1-D signal translating in time has its spectrum lying on a line passing through the origin. In the case of a band-limited signal with a bandwidth of U, the spatio-temporal transform is a line extending from $(U, -v \cdot U)$ to $(-U, v \cdot U)$ (Figure 2).



(a) (b)

(c)

Figure 1. (a) An impulse signal, (b) Translating in time, (c) Its Fourier transform in spatio-temporal frequency domain.



(a) (b)

Figure 2. (a) A band-limited signal, (b) Its Fourier transform when it is translating with speed v.

When a moving signal is sampled in time, replicas of the Fourier transform of the original signal are created on the ω (temporal frequency) axis in the transform domain, each of which are ω_s apart, where ω_s is the temporal sampling frequency. According to [1], the sampled signal is perceived the same as the continuous version, as long as the replicas lie outside the window of visibility (Figure 3). The replicas lie outside the window of visibility as long as $\omega_s \geq \omega_l + vU$, where ω_l is the edge of the window of visibility on the temporal frequency axis.

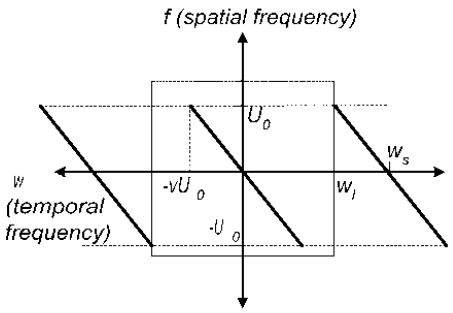
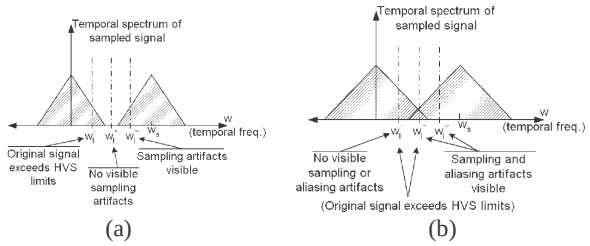


Figure 3. Temporally sampled band-limited signal, in Fourier domain.

Another consideration is the temporal aliasing due to sampling. The sampling frequency ω_s has to be at least $2 \cdot v \cdot U$ to avoid aliasing. A comparison of the aliasing and the window of visibility constraints is illustrated in Figure 4. Aliasing is a problem in computer graphics animation as well, and it is frequently handled using spatio-temporal smoothing (motion blur) [11].



(a) (b)

Figure 4. Various combinations of window of visibility (w_l) and signal bandwidth; (a) No temporal aliasing, (b) With temporal aliasing.

In the above discussion, we see the temporal bandwidth of the visual stimuli as the limiting factor on temporal sampling frequency. We stated that the temporal bandwidth of a translating 1-D signal is given by $v \cdot U$. This is better illustrated in Figure 5.

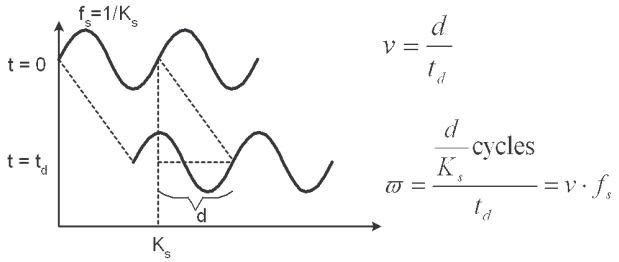


Figure 5. A translating 1-D sinusoid and the derivation of its temporal frequency.

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.

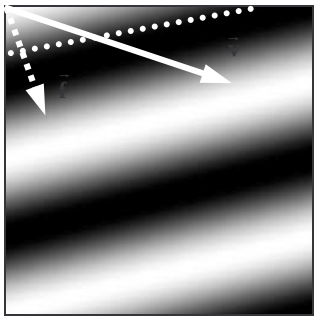


Figure 6. A 2-D sinusoid with a frequency vector **f** perpendicular to the wave front, and a motion vector **v** showing its translation velocity.

Figure 6 shows the sinusoid $\cos(2\pi\frac{1}{N}x+2\pi\frac{4}{N}y)$, where the origin is at the upper left corner, and positive y-axis is downward. Each 1-D cross-section of a 2-D sinusoid is a 1-D sinusoid. For the Figure 6, the frequency of the sinusoid along the x-axis is $f_x=\frac{1}{2}$; and the frequency along the y-axis is $f_y=2$. We represent this sinusoid with a frequency vector $\vec{\mathbf{f}}=(0.5,2)$, which points in the highest frequency direction (the gradient). Let the motion vector describing the translation of this sinusoid be given as $\vec{\mathbf{v}}=(v_x,v_y)$. Then we can show that the spatial frequency of the 1-D cross-section in the $\vec{\mathbf{v}}=(v_x,v_y)$ direction is,

$$f_v=\frac{(f_x\cdot v_x+f_y\cdot v_y)}{\sqrt{v_x^2+v_y^2}}=\frac{\vec{\mathbf{f}}\cdot\vec{\mathbf{v}}}{|\vec{\mathbf{v}}|}$$

Hence, the temporal frequency of the translating 2-D signal with spatial frequency $\vec{\mathbf{f}}$ and velocity $\vec{\mathbf{v}}$ is given by $f_v|\vec{\mathbf{v}}|=\vec{\mathbf{f}}\cdot\vec{\mathbf{v}}$. We define this product as our spatio-temporal complexity (or visual complexity) measure. In the following section, we show the computation of a compressed domain feature based on this result.

4. COMPUTATION OF THE SPATIO-TEMPORAL COMPLEXITY IN COMPRESSED VIDEO

Compressed domain algorithms and features are attractive because of significant computational, buffering, and storage gains. In many real life applications, compressed domain solutions are the only viable option. In this section, we develop a compressed domain implementation of the visual complexity feature that we have described. We consider Mpeg-1/2 compressed video, or similar compression formats where we have DCT blocks and motion compensation vectors.

4.1 Spatio-temporal Complexity From DCT

In the previous section, we showed that the temporal frequency of a translating 2-D sinusoidal grating is given by $\vec{\mathbf{f}}\cdot\vec{\mathbf{v}}$. 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 $\frac{k_y}{2}$ (Figure 7), whereas a 2-D sinusoidal grating with a frequency f_x in the x direction and f_y in the y direction is represented as $\cos(2\pi\frac{f_x}{N}x+2\pi\frac{f_y}{N}y)$.

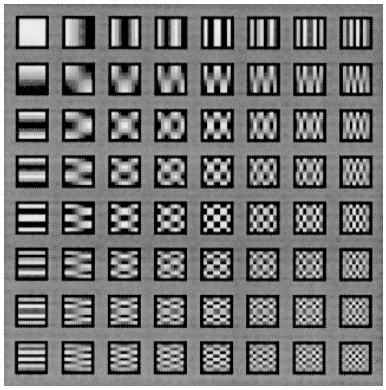


Figure 7. DCT basis images for an 8x8 block.

Using the identity

$$\cos(\ a\cdot b\ )=\frac{1}{2}\left[\cos(\ a+b\ )+\cos(\ a-b\ )\right]$$

we can write the DCT basis as;

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

Thus, each DCT basis is a superimposition 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 8). Then the temporal frequencies (or the spatio-temporal complexity) resulting from the (k_x,k_y) DCT coefficient and a motion vector $\vec{\mathbf{v}}=(v_x,v_y)$ are;

$$\omega_1 = \vec{\mathbf{f}}_1 \cdot \vec{\mathbf{v}}_1 = \frac{k_x}{2} v_x + \frac{k_y}{2} v_y \text{, and } \omega_2 = \vec{\mathbf{f}}_2 \cdot \vec{\mathbf{v}}_2 = \frac{k_x}{2} v_x - \frac{k_y}{2} v_y$$

which are in cycles-per-block units since (k_x, k_y) have that units. To convert the frequency into cycles-per-frame, we convert (k_x, k_y) into cycles-per-pixel by dividing by 8 (block size). In addition, we use the absolute values $|\omega_1|$ and $|\omega_2|$ in our computations because the sign of the frequency is irrelevant in one dimension. The $\frac{1}{2}$ factor in the DCT expansion into sum of sinusoids is also irrelevant since all the terms have the same factor. Hence the final form of the spatio-temporal complexity terms contributed by each DCT coefficient is;

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

Each DCT coefficient contributes a value equal to its energy to the bins corresponding to ω_1 and ω_2 in the spatio-temporal complexity histogram, as will be described in following sections.

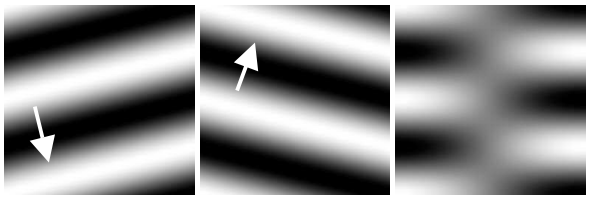


Figure 8. The two 2-D sinusoids that make up a DCT basis when summed up.

4.2 Motion Vector and DCT Estimation

Compressed domain motion compensation vectors are computed with the goal of maximizing compression efficiency and not the prediction of the real motion, making the motion vectors unreliable. Spurious vectors are common especially if the encoder is not well optimized. In order to eliminate spurious motion vectors, we first discard low-texture blocks since the block matching, which is used in finding the motion vectors, is less reliable for those blocks [12]. We implement this as thresholding of spatial bandwidth of each block, which we already compute for the visual complexity measure. Note that low-texture, i.e. low spatial bandwidth, blocks are expected to have low visual complexity, hence the risk of losing critical blocks is minimal. We then apply median filtering to further eliminate spurious motion vectors. We use interpolation to fill in the motion vector information for intra-coded (no motion vector) macroblocks.

Fitting a global motion model can also further eliminate spurious motion vectors but would also eliminate foreground object motion. However, if the application permits, global motion fitting, especially through iterated weighted least squares, can be used to increase the reliability of the motion vector field [13]. This would also eliminate the problem of intra-coded macroblocks. In the future, we want to treat background and foreground motion separately, especially in the context of human visual system's tracking of moving foreground objects.

We have the DCT coefficients for I-frames but no motion vectors. Similarly, we have the motion vectors for the P-frames, but the DCT coefficients for the motion residue only. We can compute the DCT coefficients of P frame blocks by applying motion compensation or estimate without decoding [14]. An alternative solution is by considering the motion vectors from the I-frame to the following P (or other) frame as the motion of blocks on a non-regular grid in the I-frame. Then we can interpolate the motion vector field or fit a parametric model, to obtain the motion vectors for the I-frame blocks. This is an easier and faster approach. However, foreground object motion can be lost if a parametric model is fit to the irregular motion field.

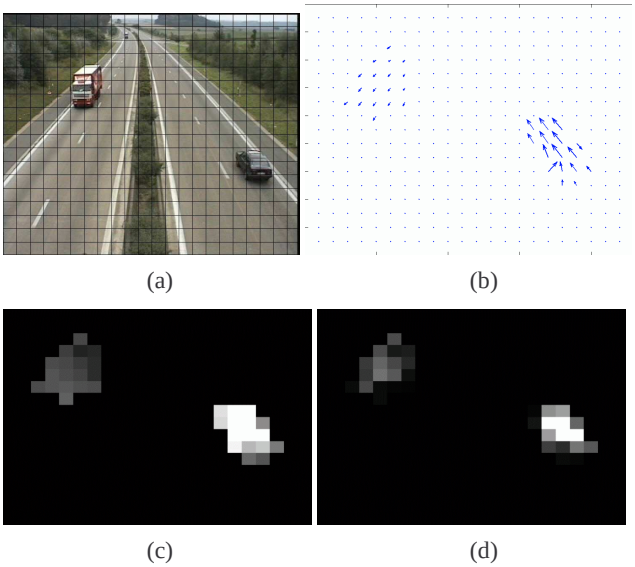
4.3 Spatio-temporal Complexity of a Video Segment

We define both a histogram-based measure and a single number measure for the visual complexity. For each macroblock, we compute the spatio-temporal complexity contribution (ω_1 and ω_2) for each DCT coefficient, and construct a histogram of the complexity distribution. We compute the complexity histogram for the frame by averaging the macroblock complexity histograms. The averaging can be performed over a number of frames for a video segment complexity as well, if required by the application. The spatio-temporal complexity histogram enables us to compute the energy that lies above a given temporal frequency. This will be used in computing the playback rate for each video frame or segment so that the quality loss is the same over all frames of the video.

A more compact measure can be derived when a histogram is too complex for the application. The average or a certain percentile can be used as a single representative figure for the spatio-temporal complexity. The spatio-temporal complexity histogram is analogous to the power spectrum, while its single number alternative is similar to a bandwidth measure.

The visual complexity measure is, in fact, an approximation of the temporal bandwidth of a video segment. Ideally, the temporal bandwidth can be computed through a 3-D FFT or DCT. However, this is impractical due to the computational complexity and the buffer requirements. The piece-wise linear motion assumption in using motion vectors allows us to estimate the temporal bandwidth easily in compressed domain.

Note that, the estimated temporal bandwidth in the form of spatio-temporal complexity may be higher than the highest possible frequency given the temporal sampling rate. This is due to a number of factors such as the error in motion vectors, the low resolution of motion vector field (block based, not optic flow), the block motion residuals, the linear motion assumption over a number of frames, etc. For example, the car in Figure 9 moves at around 10 pixels per frame, which is very large compared to its size. Indeed, the spatio-temporal complexity in that area is as high as 1.6 in some macroblocks, where 0.5 is the temporal aliasing limit. However, the spatio-temporal complexity is still a good approximation and an intuitive indicator of the visual scene complexity as it combines two important visual complexity components, the spatial detail and the motion activity level of a video frame.



(a) (b)

(c) (d)

Figure 9. (a) Frame 1284 from MPEG-7 test video speed5, (b) The motion vectors, (c) The motion activity image, (d) The spatio-temporal complexity image.

5. ADAPTIVE FAST PLAYBACK METHOD

Under the right conditions, the human visual system can see spatial resolutions up to 60 cycles/degree [15]. However, this number varies by the luminance, the contrast and the foveal location of the stimuli. Watson et. al. report spatial resolution limits of 6 to 17 cycles/degree, which reflect the imperfect lighting and contrast settings that is more likely to be found in daily life [1]. 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. The recommended viewing angle (horizontal) is about 10° for standard resolution TV and 30° for HDTV, which correspond to viewing distances of 8 and 3 screen heights, respectively (Figure 10). Since the horizontal screen resolutions are 720 (360 cycles) and 1920 (960 cycles), respectively, we have spatial resolutions around 30 cycles/degree. The VCD format has horizontal and vertical resolutions (352x240 NTSC Mpeg-1) that are almost half the DVD (720x480 NTSC Mpeg-2), and is accepted as close to VHS quality. We will take 30 cycles/degree as the high-quality spatial resolution limit (DVD), 15 cycles/degree as acceptable quality resolution (VHS) and 7 cycles/degree as low-end acceptable resolution (Watson et. al.).

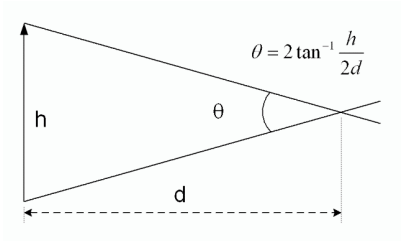


Figure 10. Conversion between angular and distance units for resolution computations.

We will take the original frame rate of the video as the visual temporal frequency limit ω_1 because it is close enough to the estimated real value, and is determined considering the human visual system. Also, it defines the highest temporal frequency allowed in the original content. Under this condition, the highest temporal frequency allowed by the window of visibility constraint is equal to the Nyquist frequency for the original playback frame rate. For example, a DCT block that has significant energy at one of the (8, n) or (m, 8) coefficients can have only 1 pixel/frame motion in that direction. In general;

$$\omega_1 \le \frac{1}{2} \text{ and } \omega_2 \le \frac{1}{2} \text{ , hence } |k_x v_x \pm k_y v_y| \le 8 ,$$

where (k_x, k_y), $1 \le k_x, k_y \le 8$, is the DCT coefficient number. This can be interpreted as an available spatial bandwidth, given the block motion. As a result, when the video playback is speeded up the motion vectors are scaled up and the allowed spatial bandwidth shrinks proportionally. Given the spatio-temporal complexity of a video segment, the maximum speed-up factor it can be played back with before temporal aliasing is,

$$f \le \frac{1}{2\omega}, \quad \omega : \text{spatio} - \text{temporal complexity} .$$

As mentioned earlier, sometimes the original spatio-temporal complexity figure is above the aliasing limit, as in Figure 9. We can still see the overall object, although we may need to slow down the video to be able to see the details. In real life, the eyes track the objects in attention, decreasing the effective speed and increasing the allowed spatial resolution at a given speed.

In cases where the video is to be played back at a rate higher than indicated by the spatio-temporal complexity, spatio-temporal filtering (motion blur) needs to be applied to avoid aliasing. In this lossy speed-up case, the spatio-temporal complexity histogram allows the computation of the energy that has to be filtered out at a given playback frame rate. Then, all the parts of the video can be speeded up so as to have the same level of loss through out the whole video. If the simpler, single number spatio-temporal complexity measure is used, video segments are speeded up inversely proportional with their spatio-temporal complexity values.

The spatio-temporal smoothing is a filtering operation in 3-D space, consisting of spatial and temporal dimensions. Temporal filtering is achieved by a weighted average of buffered frames in the MPEG decoder. The filtering removes the part of the video signal that lies outside the window of visibility, which in our case is equivalent to the aliasing limits. Since the temporal bandwidth of the video segment is the product of the spatial bandwidth and the motion, we can reduce the temporal bandwidth by spatial filtering as well as temporal smoothing. Techniques like coring allow for efficient compressed domain spatial filtering of video [16]. In applications that require low complexity, the unfiltered video can still be used though with some artifacts.

Another application dependent modification that can be employed is the smoothing and/or quantization of the spatio-temporal complexity curve for the video sequence. In certain cases the continuous change of the playback rate is not feasible or desirable. In these applications, the playback rate can be determined for a given minimum length of time (or e.g. for each shot). Further, the allowed playback rates can be limited to a set of predetermined values as in commercial video and DVD players.

We have also described a practical implementation of playback rate computation through accumulation and thresholding of the motion activity in [10], which is applicable for spatio-temporal complexity as well. Further details of the adaptive fast playback method are described in that reference.

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 11) we presented in [10]. It degenerates to motion activity when the spatial complexity of the scenes is the same for all the frames in the video. Indeed, we mentioned in [10] that adaptive fast playback using motion activity is especially successful when the background is fixed and the foreground motion is not too complex, such as in many surveillance applications. The spatio-temporal complexity extends the motion activity by introducing the scene complexity as a variable. However, in certain applications that we used motion activity for, such as sports highlights detection [8], we are primarily interested in the motion itself, hence the spatio-temporal complexity measure does not provide an advantage (See Figure 12).

The presented video skimming method can be integrated with other video content analysis and summarization methods that try to extract the semantic content. The trivial way of integration is by using adaptive fast playback as the final visualization step. Conversely, adaptive sub-sampling of frames using spatio-temporal complexity can be used as an initial data reduction step as well.

Another semantic extension we plan for the future is to introducing semantic inputs to the adaptive playback rate system. For example, the skimming system can be programmed to slow down to normal playback for a few seconds on semantic cues such as the detection of significant faces or a dialog start.

As a future improvement, we want to explore the use of a non-uniform weighting across the frame. This can be a fixed weighting, e.g. where the central parts are weighted more, as well as a dynamic weighting scheme tuned to a measure of visual attention. This would eliminate a side effect we observed in our experiments, namely, the slowing of video for unnecessary details like the clutter in the background. Another future work item to eliminate unnecessary slow down of video is the identification of ‘transitional scenes’ such as turning of a close-up head, etc. These short scenes have high visual complexity, but are not really meant to be perceived in full detail, hence the algorithm can be modified to detect those scenes and skip without slowing down.

7. ACKNOWLEDGMENTS

Our thanks to Roy Wang for his MPEG analysis software, and to Hari Kalva for bringing it to our attention.

8. REFERENCES

[1] A. Watoson, 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] M.M. Yeung and B. Liu. Efficient matching and clustering of video shots. In ICIP '95, pages 338-341,1995.

[3] D. Zhong, H. Zhang, and S.-F. Chang. Clustering methods for video browsing and annotation. In SPIE Storage and Retrieval for Image and Video Databases IV, pages 239-246,1996.

[4] A.M. Ferman and A.M. Tekalp. Efficient filtering and clustering methods for temporal video segmentation and visual summarization. J. Vis. Commun. & lmage Rep., 9:336-351, 1998.

[5] D. DeMenthon, V. Kobla and D. Doermann, "Video Summarization by Curve Simplification", ACM Multimedia 98, pp. 211-218, September 1998.

[6] A. Divakaran, R. Radhakrishnan, and K.A. Peker, "Motion Activity based extraction of key frames from video shots," Proc. IEEE Int'l Conf. on Image Processing, Rochester, NY, Sept. 2002.

[7] 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.

[8] K.A. Peker, R. Cabasson, and A. Divakaran, "Rapid generation of sports video highlights using the MPEG-7 motion activity descriptor," Proc. SPIE Conf. on Storage and Retrieval for Multimedia Databases, San Jose, CA. Jan 2002.

[9] A. Divakaran, R. Regunathan, and K. A. Peker, "Video summarization using descriptors of motion activity," Journal of Electronic Imaging, vol. 10, no. 4, Oct. 2001

[10] K. A. Peker, A. Divakaran and H. Sun, “Constant pace skimming and temporal sub-sampling of video using motion activity,” Proc. IEEE Int’l Conf. on Image Processing, Thessaloniki, Greece Oct. 2001.

[11] K. Sung, A. Pearce, C. Wang, “Spatial-temporal Antialiasing,” IEEE Trans. Visualization and Computer Graphics, Vol.8 No.2, pp. 144-153, April 2002.

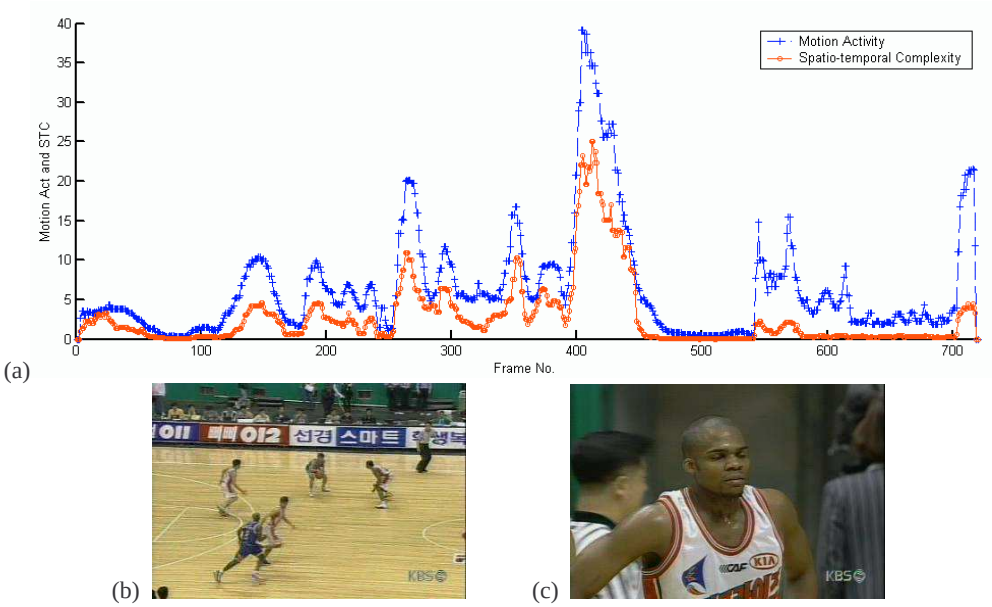
[12] M. Pilu, "Motion re-estimation from raw MPEG vectors with applications to image mosaicing", SPIE Electronic Imaging Conference, Photonics West, San Jose (CA), Jan 1998.

[13] R. Wang, T. Huang “Fast Camera Analysis in MPEG Domain”, Int. Conf. Image Processing (ICIP) 1999.

[14] S.-F. Chang and D. G. Messerschmit, “Manipulation and compositing of MC-DCT compressed video”, IEEE Journal on Selected Areas in Communications, Special Issue on Intelligent Signal Processing, Jan. 1995, pp. 1-11.

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

[16] P.M.B. van Roosmalen, R.L Lagendijk, J. Biemond, “Embedded Coring in MPEG Video Compression,” IEEE Trans. Circuits and Systems for Video Tech., Vol.12 No.3, pp. 205-211, March 2002.



(a)

(b)

(c)

Figure 11. (a) Motion activity and spatio-temporal complexity (STC) for a basketball video segment (MPEG7 testset). The two measures are similar except the last part, which is a close up on a player. STC is lower here because the images are larger with less detail compared to wide shots. (b) Frame 300, camera pan. (c) Frame 600, close-up on player.

Frame 500 Frame 1100 Frame 2220 Frame 2300

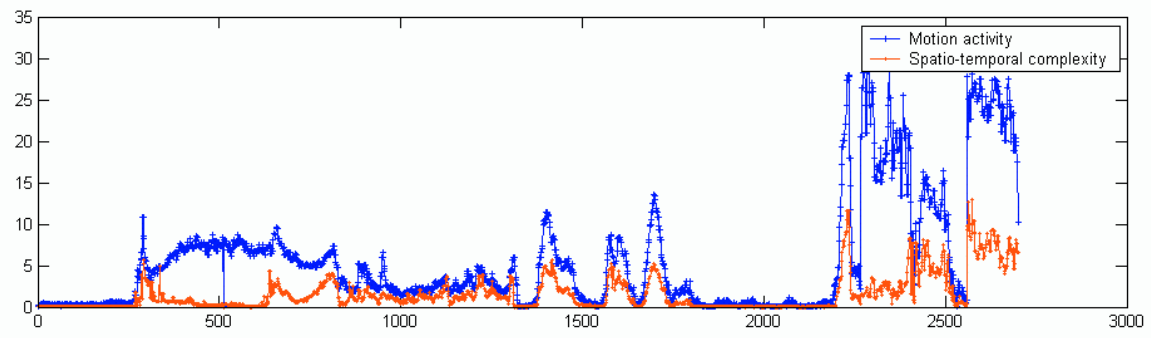


Figure 12. Motion activity and spatio-temporal complexity (STC) for a golf video segment (MPEG7 testset). STC is lower at the last part (frame 2300) where the camera tracks the ball in the air, and around frame 500 where there is camera motion over a smooth green field. STC overshoots when there are trees and bushes in the background during a pan around frame 2200.