Complexity Analysis of Video Frames by Corresponding Audio Features

SeungHo Shin and TaeYong Kim
GSAIM, Chung-Ang University, Seoul, Korea

Keywords: Video Complexity Analysis, Video Indexing, Audio Indexing, Rate-control.

Abstract: In this paper, we propose a method to estimate the video complexity by using audio features based on human synesthesia factors. By analyzing the features of audio segments related to video frames, we initially estimate the complexity of the video frames and can improve the performance of video compression. The effectiveness of proposed method is verified by applying it to an actual H.264/AVC Rate-Control.

1 INTRODUCTION

It is essential to detect video complexity with high accuracy to improve the compression performance by reducing unnecessary processing in video coding. In this aspect, a video complexity analysis with audio features has great advantages for reducing computation and simplifying the algorithm compared to the analysis using visual information only. For instance, we can search combat scenes easily by detecting frames that include gunfire or explosive sounds in a video. However, it is hard to detect these scenes by visual analysis, such as color differences, histograms between frames, object detection, motion estimation, and other various features. Therefore, in the analysis and understanding of video content, if we simultaneously use the auditory information as an auxiliary element for the visual information, it will improve the accuracy of complexity analysis.

In this paper, we propose a novel method named "Content-based Video Complexity Analysis (CVCA)" that estimates the temporal and spatial complexity based on human synesthesia factors by analyzing the correlations between the video and audio presented in moving pictures. The most important characteristic of the CVCA is the use of the variations of audio signals to analyze the complexity of a video. The effectiveness of this method is verified by applying it to an actual H.264/AVC Rate-Control.

2 CORRELATION BETWEEN VIDEO AND AUDIO FEATURES

The visual media that has a storyline is called "synthetic art" and is communicated to our eyes and ears with the video's visual information and the audio's auditory information (Li, 2004). The audio elements related to the video are able to be classified into three factors: dialog between actors, sound effects from the surrounding objects, and background sounds for scene enhancement (Lu, 2002).

1) In the scene where actors make conversation normally, the voice tone is maintained steadily and the audio signals are regular and natural. However, when the actors are arguing over something, the voice tone becomes rough, causing the audio signals to be irregular (Pinquier, 2002). In order to express the confrontational scenes effectively in video, quick camera-moving switches between actors, and the actor's motion is also increased with exaggerated gestures to express agitated emotions.

2) The environment and place represented in the scenes are variously changed according to the storyline. The sound effects are dependent on the given environment and place. Viewers can roughly recognize situations and scenes via sound effects. The most representative case would be a combat scene including the gunfire and explosive sounds.

3) The background sounds are also used to more effectively express the video scenes. The tempo and rhythm of background sounds are one of the most considered elements in the video editing works. In

Shin S. and Kim T..
Complexity Analysis of Video Frames by Corresponding Audio Features.
DOI: 10.5220/0003982001110114
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications and Wireless Information Networks and Systems* (SIGMAP-2012), pages 111-114
ISBN: 978-989-8565-25-9

addition, the loud and abrupt sounds are often used to emphasize the scene change.

In order to analyze the correlation between audio and video in actual moving pictures, we compute the correlation to extract the complexity of a video frame in accordance with the change of audio signals. In a video frame where the amplitude of audio signal on the timeline varies greatly, the difference between the video frames where audio variation happened is calculated to obtain the changes of video complexity.

To measure the strength of the linear relationship between audio and video, we used the Pearson's correlation in Eq. (1).

$$r = \frac{\sum_{i=0}^{n} (A_i - \overline{A})\,(V_i - \overline{V})}{\sqrt{\sum_{i=1}^{n} (A_i - \overline{A})^2}\,\sqrt{\sum_{i=1}^{n} (V_i - \overline{V})^2}}\ , \qquad (1)$$

where r is the correlation coefficient, A is the variable converted into audio values and V is the combination of visual features as $V = (V_{motion} + V_{itensity})$. $\overline{A}$ is the mean value of audio A and $\overline{V}$ is the mean value of video V. V_{motion} is the motion vector ratio of video frames, and $V_{intensity}$ is the color intensity difference between current and previous frames.

Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect positive linear relationship between variables, -1 is a perfect negative correlation, and 0 is no correlation. Generally, if the correlation is larger than absolute 0.5, there is able to be a valid correlation between two variables.

As shown in Fig. 1, the result shows that the correlation is higher in the genres that have lots of dynamic scenes such SF, fantasy, action, war and horror. On the other hands, correlation is low in the genres that have scenes of stillness such as educational materials and talk shows.

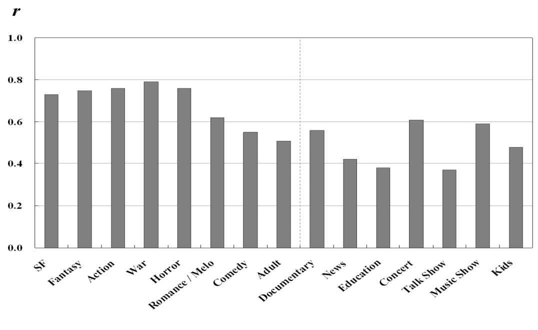

Figure 1: Correlation analysis of video contents.

3 RATE-CONTROL BASED ON CVCA

When the available bandwidth and the compression rate are constant, the "Rate-Control (RC)" becomes the most important factor for the improvement of objective and subjective quality in video compression (Ma, 2002). In order to transmit the compressed data safely, available bits have to be allocated according to the video complexity. In this section, we propose the "Content-based Video Complexity Analysis (CVCA)" that uses audio features, and suggest a method that improves the rate-control efficiency.

3.1 Estimation of Video Complexity by Audio Feature Parameter (AFP)

The time-domain characteristics of an audio segment are composed of six features: the standard deviations of the frame energy, the silence ratio, the zero-crossing ratio, the volume root mean square, the volume dynamic ratio, and the total energy. These audio features reflect important characteristics related with video features (Lu, 2002); (Pinquier, 2002). In this paper, we use the total energy for audio feature parameter (*AFP*) and it can be defined as follows:

$$AFP_n = 10 \log\left(\frac{1}{N} \sum_{i=1}^{N} x^2 \right), \qquad (2)$$

where x is the amplitude of an audio signal sample. To extract the *AFP*, 5ms audio samples are combined into a 25ms unit and then 20 units are grouped together as an audio frame, AFP_n.

The *AFP* difference for each audio frame is measured by:

$$AFP_{diff} = \left| \frac{\sum_{i=0}^{N} AFP_n^i - \sum_{i=0}^{N} AFP_{n-1}^i}{\sum_{i=0}^{N} AFP_n^i} \right| \times 100 . \qquad (3)$$

Finally, the strength of the segment complexity (AFP_r) is estimated by the AFP_{diff} and *AFP* as shown in Table 1.

3.2 Video Frame Complexity Analysis by Video Feature Parameter (VFP)

After estimating the complexity of a video frame by AFP, the actual analysis of the complexity is performed by the video feature parameter (*VFP*).

Table 1: Complexity decision for audio segments.

Rules	Segment complexity
if $(AFP_{diff} \geq T_{\alpha 0})$ *and* $(AFP_n \geq T_{\beta 0})$	$AFP_r = 1.0$
if $(AFP_{diff} \geq T_{\alpha 0})$ *or* $(AFP_n \geq T_{\beta 0})$	$AFP_r = 0.75$
if $(T_{\alpha 0} > AFP_{diff} \geq T_{\alpha 1})$	$AFP_r = 0.5$
if $(T_{\beta 0} > AFP_n \geq T_{\beta 1})$	$AFP_r = 0.25$
else	*none*

where T_α and T_β are thresholds for mode classification.

Generally in the video frames that represent sudden changes in audio signals, there are increased scene changes and motion activities. However, it is necessary to analyze video frames in order to prevent the misinterpretation of exceptional cases.

The *VFP* consists of two elements: VF_t which represents the temporal complexity and VF_s which represents the spatial complexity of video frames. The VF_t is determined by the motion vector difference and sum-of-absolute difference between current and previous frames in Eq.(4).

$$VF_t = \frac{1}{M}\sum_M |MV_n - MV_{n-1}| + \frac{1}{N}\sum_N |f_n - f_{n-1}|, \qquad (4)$$

where M is the total number of macro blocks, N is the number of pixels, and MV represents a motion vector for a pixel f in a frame.

The VF_s is determined by the spatial derivative using Sobel operator for real-time coding and low implementation complexity.

$$VF_s = \frac{1}{N}\sum_N \sqrt{G_x^2 + G_y^2} \ , \qquad (5)$$

where G_x and G_y are derivatives to horizontal and vertical directions, respectively.

3.3 Rate-control based on the CVCA

In the segments that have the sudden increase of video complexity compared to the normal segments, instantaneous deterioration of video quality may occur. Thus, in order to prevent such video quality degradation, the bit allocation scheme is applied to each frame according to *AFP* and *VFP*.

The quantization parameter (QP) for each frame is recalculated by the AFP and VFP, which is defined as:

$$Q^* = \frac{Q}{AFP_r} \left| F_{type} \cdot VF_t + (1 - F_{type}) \cdot VF_s \right| \qquad (6)$$

where Q is the original QP, VF_t is the temporal complexity of the current frame, and VF_s is the spatial complexity of the current frame. F_{type} is a weight for frame types. In an intra-frame case, because it does not remove the temporal redundancy, the amount of bits for this frame is proportional to the spatial complexity. Therefore, the intra-frame coding is affected only by VF_s. On the contrary, the amount of bits is proportional to the temporal complexity in the inter-frame case. The target bits by Q^* are allocated in either intra or inter frame by applying Rate-Quantization (R-Q) model (Kwon, 2003).

4 EXPERIMENTAL RESULTS

In order to measure the performance of the proposed RC method, we experimented by applying it to the H.264/AVC rate-control. The CVCA analyzer transfers the complexity information to an encoder, then the encoder controls the bit-rate for each video frame based on this information. To control the bit-rate in a limited bandwidth, the encoder partly saves the unnecessary bits of the normal segments and assigns these saved bits to the complex segments.

The video clips in this experiment consist of movies, dramas, and animations that have narrative structures. Each video clip is 30 minutes in duration and includes simple and complex scenes. The format is QVGA, and the luma and chroma components are sampled at 4:2:0. The PSNR (Peak Signal to Noise Ratio) is used to evaluate the objective quality of the video, and we also evaluate the subjective quality. For demonstrating the performance of the proposed RC, we compare it to the original H.264 RC.

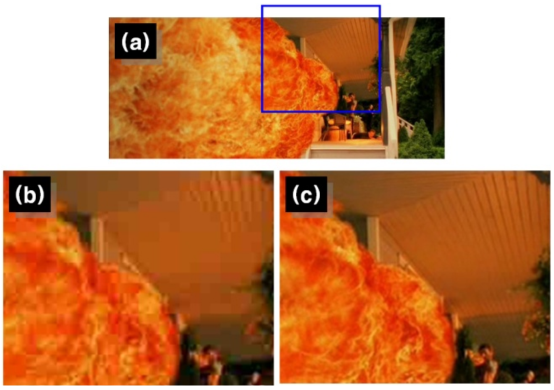

Figure 2: The comparison of the subjective quality (Seq.1); (a) the original image, (b) the close-up view of the reconstructed frame using the original RC, and (c) the close-up view of the reconstructed frame with the proposed RC.

Fig. 2 and Table 2 show that the proposed RC significantly outperforms the original RC in the segments having abrupt changes. In the segments where the *AFP* is detected, the *VFP* is generally increased by object motions, camera-moving or scene changes. As a result, the PSNR is decreased in the original RC, but the proposed RC is more stable than original RC. In the normal segments where the AFP is not detected, there is a little PSNR decrease.

Table 2: PSNR comparison between original and proposed RC in complex segments of videos.

Sequence	Average PSNR(dB) (Original RC)	Average PSNR(dB) (Proposed RC)	Difference (dB)
Seq. 1	27.1	30.6	3.5
Seq. 2	28.3	31.2	2.9
Seq. 3	30.6	33.1	2.5

Because the rate-control based on the CVCA is able to use the prior estimation of complexity for a video sequence, it is possible to improve the video quality in the compression.

5 CONCLUSIONS

In this paper, we propose a method "Content-based Video Complexity Analysis (CVCA)" to enhance the video complexity analysis with audio features, which improves the rate-control efficiency in the compression. The correlation between audio and video in moving pictures is demonstrated, the complex segments in a video sequence are estimated initially by audio features, and actual temporal and spatial complexities for the estimated segments are analyzed by visual features. Then the bit allocation scheme is applied to each frame, and the rate-control efficiency is improved.

ACKNOWLEDGEMENTS

This research was supported by the MKE under the HNRC-ITRC support program supervised by the NIPA (NIPA-2010-C1090-1011-0010) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0787).

REFERENCES

Kwon, J. C., Lee, M. J., Kim, J. K., 2003. Practical R-Q and D-Q Models for Video Rate Control*, IEICE Trans. On Communications*, vol.E86-B, no.1.

Li, Y., Narayanan, S. S., Kuo, C. C. J., 2004. Content-Based Movie Analysis and Indexing Based on Audio-Visual Cues, *IEEE Trans. Circuits Syst. Video Technol*., vol.14, no.8, pp.1073-1085.

Lu, L., Zhang, H. J., Jiang, H., 2002. Content Analysis for Audio Classification and Segmentation, *IEEE Trans. On Speech and Audio Proc., vol.10, no.7, pp.504-516.*

Ma, S., Gao, W., Lu, Y., 2002, *Rate control on JVT standard,* JVT-D030, pp.22-26.

Pinquier, J., Rouas, J., 2002. Robust Speech/music classification in audio documents, *International Conference on Spoken Language Processing, Denver,* USA, vol.3 pp.2005-2008.