Video Content-Based Bit Rate Estimation Scheme for Transcoding in IPTV Services

Hye Jeong Cho[1] , Chae-Bong Sohn[2] and Seoung-Jun Oh[3]
[1] AV Research and Development Laboratory, ARION Technology Inc.,
Gyeonggi-Do, Korea
[e-mail: innocent11@gmail.com]
[2] Department of Electronics and Communications Enginering, Kwangwoon University
Seoul, Korea
[e-mail: cbsohn@kw.ac.kr]
[3] Department of Electronic Enginering, Kwangwoon University
Seoul, Korea
[e-mail: sjoh@kw.ac.kr]
*Corresponding author: Seoung-Jun Oh

Received October 31, 2013; revised January 20, 2014; accepted January 4, 2013; published February 9, 2014

Abstract

In this paper, a new bit rate estimation scheme is proposed to determine the bit rate for each subclass in an MPEG-2 TS to H.264/AVC transcoder after dividing an input MPEG-2 TS sequence into several subclasses. Video format transcoding in conventional IPTV and Smart TV services is a time-consuming process since the input sequence should be fully transcoded several times with different bit-rates to decide the bit-rate suitable for a service. The proposed scheme can automatically decide the bit-rate for the transcoded video sequence in those services which can be stored on a video streaming server as small as possible without losing any subject quality loss. In the proposed scheme, an input sequence to the transcoder is sub-classified by hierarchical clustering using a parameter value extracted from each frame. The candidate frames of each subclass are used to estimate the bit rate using a statistical analysis and a mathematical model. Experimental results show that the proposed scheme reduces the bit rate by, on an average approximately 52% in low-complexity video and 6% in high-complexity video with negligible degradation in subjective quality.

Keywords: IPTV, Smart TV, video transcoder, automatic bit rate estimation, QoE, candidate frame, hierarchical clustering

This work was supported by the Research Grant of Kwangwoon University in 2012.

http://dx.doi.org/10.3837/tiis.2014.03.019

1. Introduction

Broadcasting and communications convergence services use limited networks to deliver IPTV, Smart TV, and other Internet services to consumers. The compression technology of serviced video is applied according to two business models: Managed Network and Open Internet. IPTV service providers deliver the H.264/AVC video content through the Managed Network. In the Open Internet, the service is delivered over the public Internet and should enable access to video content not only from TV sets but also from other home devices, such as portable multimedia players and laptop computers. The scalable video coding (SVC) technology enables the system to consider the available bandwidth for other devices. The fully implemented SVC, however, also comes with some increase in complexity and bit rate for the same fidelity as compared with single-layer coding [1]. A further study is needed on how to best control the SVC rate according to the network resource availability [2]. Most IPTV services are focused on delivering high-resolution/high-quality video over the Managed Network, with supporting quality of service (QoS).

The MPEG-2 standard has been widely deployed in video distribution infrastructures, such as cable and satellite networks, as well as in several consumer applications, such as DVDs and DVRs. The H.264/AVC standard is used in many video streaming services limited by the network bandwidth and offers a significant reduction in the bit rate over earlier standards-based technologies such as MPEG-2 (65%) and MPEG-4 (40-50%) [3] [4]. The standard achieves better performance in terms of both the peak signal to noise ratio (PSNR) and visual quality at the same bit rate as compared with prior video coding standards.

In video streaming services with IPTV and Smart TV, a video transcoder is necessary to leverage the compression efficiency offered by H.264/AVC with broadcast quality content produced in the MPEG-2 format. To service video content over the Managed Network for users, **Fig. 1** shows the process used for video content transmission.

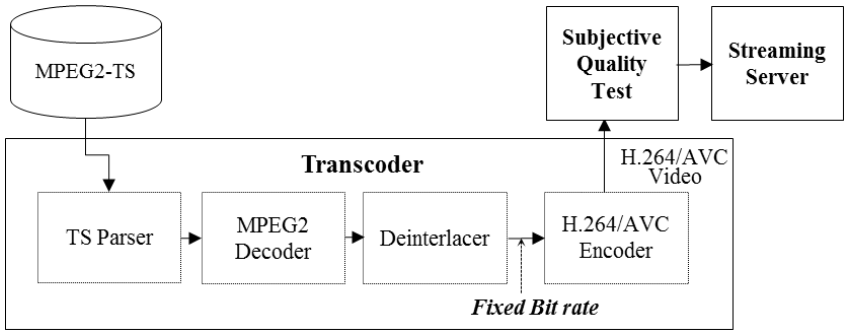


Fig. 1. Video content transcoding process in an IPTV service.

In the transcoder, the input video is decoded by MPEG-2 and re-encoded by H.264/AVC at a fixed bit rate. After performing the validation of subjective quality, the video content is stored on a video streaming server and then serviced to users with varied, engaging content via a streaming server [5]. The encoded video content is usually delivered through constant bit rate (CBR) channels. The bit rate channels needed for SDTV and HDTV video can be as high as 2–3Mbps and 10–12Mbps, respectively. Each item of video content on a CBR channel does not take into account the content's characteristics because it is encoded by two different fixed bit rates; however, the serviced video content varies from low-complexity video to

high-complexity video. The former can be encoded with a bit rate less than the fixed bit rate, without degradation in subjective quality. In other words, the conventional scheme based on a fixed bit rate causes bandwidth loss and requires a huge amount of storage space on a streaming server. When the open IPTV service is activated later, IPTV service providers can deliver the content, which, unlike specific companies' customized content, is a network resource that anyone can access. In order to deliver a considerable amount of content on a CBR channel, it is important to select an efficient bit rate.

Solving this problem requires a scheme capable of finding an appropriate bit rate for video content while maintaining a subjective quality equivalent to that of a scheme that uses a fixed bit rate. Employing this scheme requires determining a bit rate for video content prior to encoding it. A video transcoder can provide an additional controller that can also estimate the bit rate. A simple technique to estimate the video content's bit rate is to vary the bit rate step in the H.264/AVC encoder part of the transcoder. The visual quality should be verified at each encoding pass. Even though this method can provide an accurate bit rate, it is a very time-consuming process. The time required to estimate the bit rate should be minimized to meet the video streaming service requirements.

In this paper, a scheme is proposed for automatically estimating the bit rate of each subclass without the repeated full encoding and subjective quality test. Using parameters, the video content is divided into several segments. To estimate the bit rate of each segment, candidate frames are extracted, which include intra-frames that require a high number of bits. Finally, the bit rate of each segment is estimated by statistical analysis and a mathematical model based on a given target quality. The remainder of this paper is organized as follows. Section II explains the analysis of video content with respect to the quality and bit rate. Section III proposes a bit rate estimation scheme for unsupervised segmentation using the frame complexity of video content. Then, the experimental results and conclusions are presented in Sections IV and V, respectively.

2. Analysis of the Quality and Bit Rate of Video Content

The purpose of this analysis is to examine the human perceived quality corresponding to the bit rates of a video. The subjective quality of the H.264/AVC encoded video is evaluated, in which a low-complexity content category such as "lecture" is coded at bit rates from 1.0 to 2.5Mbps. The evaluation is performed using the double-stimulus continuous quality scale (DSCQS) method of ITU-R Rec. BT.500-7 [6]. All the coded stimuli are rated by each of the five viewers. General conclusions were based on the quality ratings of the presented stimuli. The main idea of measuring the DSCQS score is to determine the differential mean opinion score (DMOS) between the reference encoded at 2.5Mbps and the test sequences averaged by all the viewers. A DMOS value, *dMOS*, is defined as follows:

$$dMOS = MOSr - MOSp \tag{1}$$

where *MOSr* is the MOS of the reference sequence encoded at 2.5Mbps, and *MOSp* is the MOS of the test sequence encoded below 2.5Mbps. The task is to assess the degradation of the test sequence with respect to the reference sequence. If *dMOS* is near "0", then the test sequence is similar to the reference sequence. **Fig. 2** shows the result of the average of all *dMOS*'s in a low-complexity video. The quality degradation determined by the video encoded

bit rate was, on an average, 1.4Mbps. Therefore, the low-complexity video can encode a bit rate lower than 2.5Mbps, with negligible degradation of subjective quality.

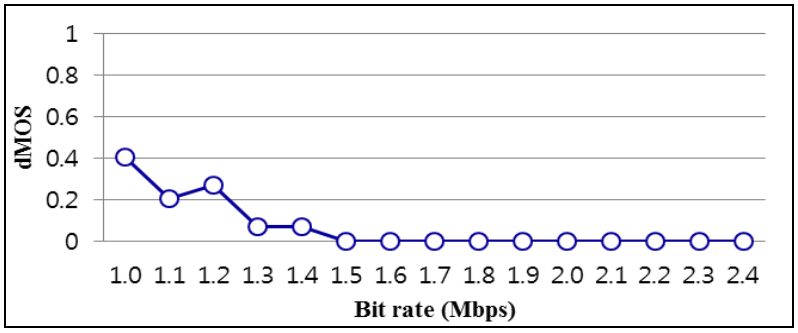


Fig. 2. Result of quality evaluation.

Further, the difference between the variable bit rate (VBR) at QP 22 and the CBR at 2.5Mbps is analyzed for the test sequence. As shown in **Fig. 3**, some video content can be encoded at a lower bit rate than at the fixed bit rate. Video content can be divided into two or three subclasses in terms of the quality of experience (QoE). It can also be delivered using more than one bit rate according to subclasses in a CBR channel.

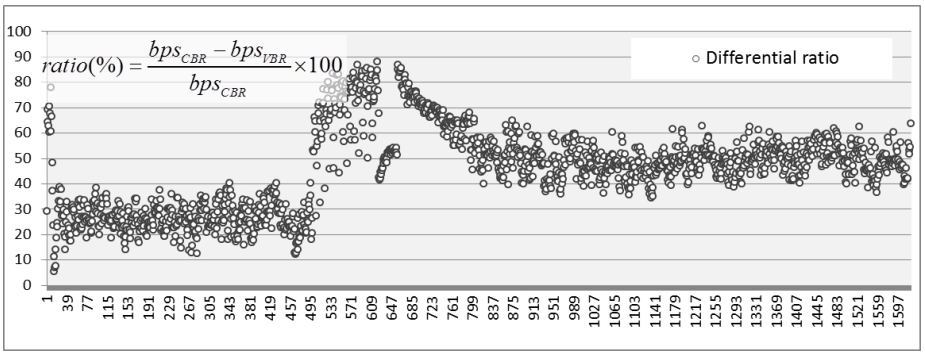


Fig. 3. The differential ratio between VBR and CBR.

3. Proposed Scheme

In this section, a bit rate estimation scheme is proposed that reduces the bit rate while maintaining the target quality in video streaming services limited by the network bandwidth. **Fig. 4** shows a block diagram of the proposed scheme. Given an input sequence as MPEG-2 TS, the TS parser is used to gather MPEG-2 video data and their data is decompressed by MPEG-2 decoder. Deinterlacer performs deinterlacing interlaced video frames to progressive video frames because a common way to compress video is to interlace it. Using those parameters, the frames of video can be divided into several segments. To estimate the bit rate of each segment, candidate frames are extracted, which includes intra-frames that require a large number of bits. Finally, the bit rate of each segment is estimated by statistical analysis and a mathematical model based on the target quality. The input video is re-encoded by

H.264/AVC at estimated bit rate. After performing the validation of subjective quality, the video content is stored on a video streaming server.

The proposed scheme differs from the conventional scheme in that it employs a bit rate estimator. Because the proposed scheme does not encode full frames of video content, it is very important to determine parameters that can serve to indirectly measure a frame's bits.

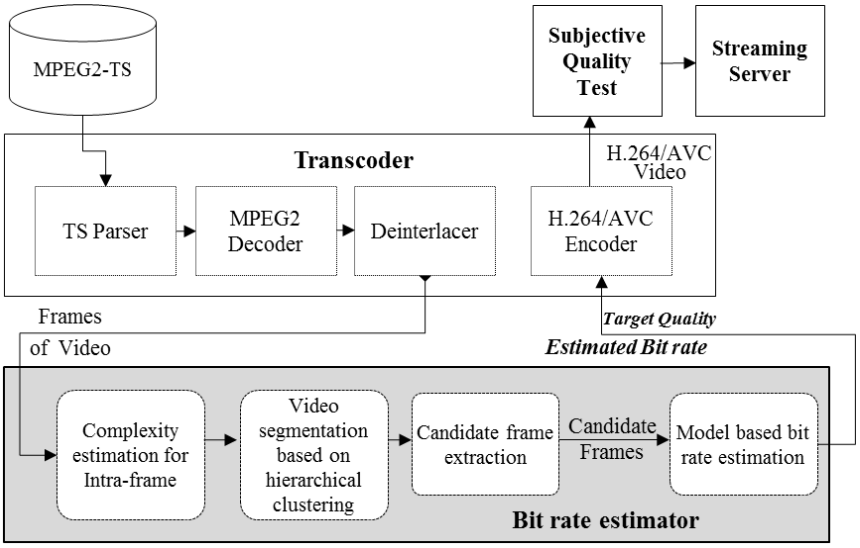


Fig. 4. Block diagram of the proposed scheme.

3.1 Frame Complexity Estimation for an Intra-frame

Some content complexity measurements for coding still images can be obtained without pre-encoding by using variance, edge, and gradient methods [7]. From the deviation of each macroblock (MB), the complexity can also be determined [8]. In the gradient-based method, the computation for calculating the gradient is low, and the output bit rate of each intra-frame is highly correlated [9]. These properties are highly desirable for measuring the complexity of an intra-frame. In addition to the gradient information, the histograms of luminance and chrominance pixel values are also very useful when combined with the gradient to represent the content complexity.

Given the arbitrary *s*th test sequence $\mathbf{Q_s}$, the set contains a number of groups of pictures (GOPs) specified in the order in which the intra- and inter-frames are arranged:

$$\mathbf{Q_s} = \{\{Q_{s(1,1)},...,Q_{s(1,N)}\},...,\{Q_{s(M,1)},...,Q_{s(M,N)}\}\} \tag{2}$$

where M is the total number of GOPs, and N is the number of frames in a GOP. $Q_{s(i,j)}$ denotes the jth frame of the ith GOP. Our objective is to measure the intra-frame complexity in $\mathbf{Q_s}$. In order to measure the frame complexity, the complexity measurement defined in [10], FC_{intra}, is used. The value of FC_{intra} for $Q_{s(i,j)} \in \mathbf{Q_s}$, CC($Q_{s(i,j)}$), can be computed by (3).

$$CC(Q_{s(i,1)}) = Grad_{s,i} \times SOH_{s,i}$$

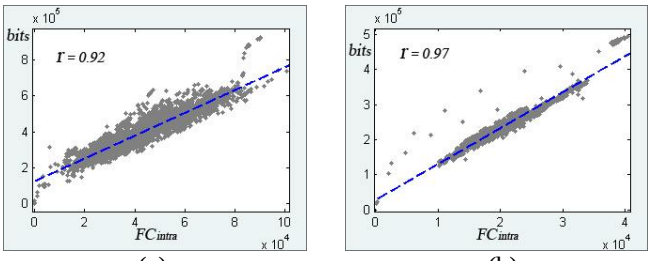
where

$$\begin{aligned} Grad_{s,i} = & \sum_{x=0}^{K_Y-2}\sum_{y=0}^{L_Y-2}(|Y_{s,i(x,y)} - Y_{s,i(x,y+1)}| + |Y_{s,i(x,y)} - Y_{s,i(x+1,y)}|)/(K_Y L_Y) \\ & + \sum_{x=0}^{K_U-2}\sum_{y=0}^{L_U-2}(|U_{s,i(x,y)} - U_{s,i(x,y+1)}| + |U_{s,i(x,y)} - U_{s,i(x+1,y)}|)/(K_U L_U) \\ & + \sum_{x=0}^{K_V-2}\sum_{y=0}^{L_V-2}(|V_{s,i(x,y)} - V_{s,i(x,y+1)}| + |V_{s,i(x,y)} - V_{s,i(x+1,y)}|)/(K_V L_V), \end{aligned} \tag{3}$$

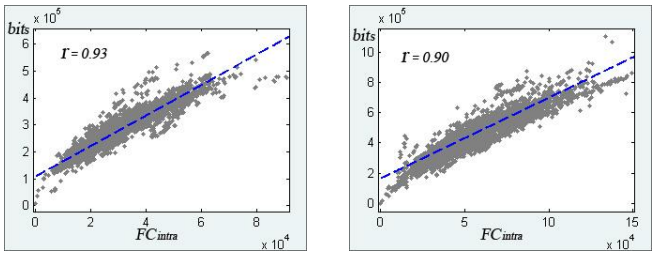
$$SOH_{s,i} = \sum_{l=0}^{255}(\log_2 HY_{s,i}[l] + \log_2 HU_{s,i}[l] + \log_2 HV_{s,i}[l])$$

In (3), $Grad_{s,i}$ and $SOH_{s,i}$ are the gradient and the statistic, respectively, of the histogram information of the ith intra-frame. $Y_{s,i(x,y)}$ is the luminance value of pixel (x, y) in the ith frame. $U_{s,i(x,y)}$ and $V_{s,i(x,y)}$ are the corresponding chrominance values. $K_Y L_Y$, $K_U L_U$, and $K_V L_V$ are the sizes of the Y-, U-, and V-frames in $Q_{s(i,1)}$. $HY_{s,i}[l]$ is the histogram of the luminance level l, and $HU_{s,i}[l]$ and $HV_{s,i}[l]$ are the histograms corresponding to the chrominance level l.

To investigate the relationship between the actual number of encoded bits and FC_{intra}, various test sequences were extensively encoded using the intra-coding mode under constant quantization parameters (QPs), and both the number of encoded bits and the FC_{intra} for each frame were recorded. **Fig. 5** shows the scatter plots of the number of bits versus FC_{intra} at different QPs in our test content, where each dot represents a frame. **Fig. 5** also shows the accuracy of the linear approximations (as blue dotted lines) by plotting the correlation coefficient r, which is an indicator of how closely the approximated linear relationship represents the actual data. The value of r lies between -1 and 1. For the test sequences, the value of r between the number of bits and FC_{intra} is, on an average, 0.93. When the value of r is at or near 1, the approximated linear relationship is the most reliable. Therefore, it is clear that a linear relationship exists in our test sequences with different slopes, and (3) can be used accurately to estimate the number of bits for intra-frames.



(a) (b)



(c) (d)

Fig. 5. Scatter plots of the number of encoded bits versus FCintra: (a) Documentary, (b) Lecture, (c) Religion, and (d) Sports.

3.2 Hierarchical Clustering-Based Video Sub-classification

Each of the subclasses—clusters, or groups of patterns of FC_{intra} —has a similar number of bits. The classifier for FC_{intra} is designed by hierarchical clustering with Bayesian decision theory [11].

Consider a sequence T containing n samples and c clusters. To conduct agglomerative hierarchical clustering for FC_{intra}, the number of initial clusters, n, is determined by analyzing the temporal characteristic between frames. The scaled-invariant feature transform (SIFT) is sequentially applied to detect stable frames among temporal frames [12]. Let $T(x,y,t)$ be the ordinal signature of the (x,y)th block of the tth frame in T. $G_\sigma(x,y,t)$ defines a 3×3×3 Gaussian kernel with standard deviation σ as follows:

$$G_\sigma(x,y,t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2+y^2+t^2)/2\sigma^2} \tag{4}$$

A 3×3×3 difference-of-Gaussian (DoG) kernel [13] is derived by computing the difference between two Gaussian kernels as follows:

$$\mathrm{DoG}_s(x,y,t) = G_{k^{s+1}\cdot\sigma}(x,y,t) - G_{k^s\cdot\sigma}(x,y,t) \tag{5}$$

where $k > 1$ is a multiplicative factor, and $s = 1,2,...,$ is the scale of the DoG kernel. Then, the DoG kernel sliding over T is used to generate a vector ψ by the convolution operation as follows:

$$\psi(t) = \sum_{t'=t-1}^{t+1} \sum_{x=1}^{3} \sum_{y=1}^{3} T(x,y,t') \cdot \mathrm{DoG}_s(x,y,t') \tag{6}$$

for $t = 1,...,m$. If the tth element in ψ is a local extreme, it is considered to be a key frame in T. In this paper, the parameters are set to $\sigma = 1.8$, $k = \sqrt{2}$, and $s = 3$. A sequence consists of the static subclass ω_0 and dynamic subclass ω_1 divided by distribution of ψ. The two subclasses are defined as follows:

$$\omega_0 : \psi(t) = \psi(t-1)$$
$$\omega_1 : \psi(t) \neq \psi(t-1), \tag{7}$$

where ω_0 denotes the same value between the tth element and (t-1)th element in ψ, whereas ω_1 denotes the different value between them. The number of initial clusters n is decided by the intervals of successive ω_0's and the number of ω_1's. **Fig. 6** shows the number of initial clusters in a sequence.

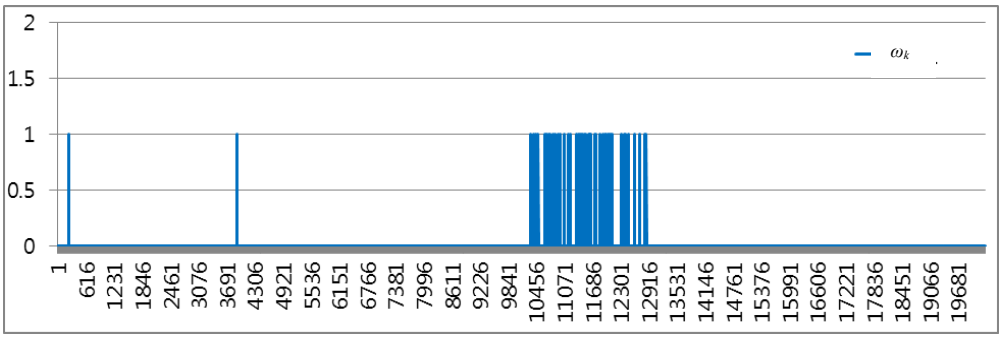


Fig. 6. Examples of the number of initial clusters

To show ω_0 and ω_1 for the distribution of frame variations, the lines in the figure denote 0 and 1 for ω_0 and ω_1, respectively. The number of initial clusters in a sequence is finally 71 as shown in **Fig. 6**. Each cluster center is the average of FC_{intra}'s in ω_0 and an FC_{intra} in ω_1, respectively. The measure of the distance between two clusters uses the Euclidean metric [14].

Given two clusters, whether they are in the same subclass or not is decided by the Bayesian decision theory. This approach is based on quantifying the trade-offs between various classification decisions using probability and the costs that accompany such decisions. It makes the assumption that the decision problem is posed in probabilistic terms and that all of the relevant probability values are known. More generally, assume that there is a prior probability $P(\omega_k)$ of each subclass k. These prior probabilities reflect prior knowledge of how likely it is that the static or dynamic subclass can be obtained before a sequence actually appears. The difference between the representative FC_{intra}'s in the two clusters is measured. Its value x is considered to be a random variable whose distribution depends on the class and is expressed as $p(x|\omega_k)$. To determine the subclass of a cluster, the following decision rule is used: decide ω_0 if $P(\omega_0|x) > P(\omega_1|x)$; otherwise decide ω_1. The decision rule can be expressed as follows:

$$P(\omega_1 \mid x) \underset{\omega_0}{\overset{\omega_1}{\gtrless}} P(\omega_0 \mid x) \tag{8}$$

Suppose that both the prior probabilities $P(\omega_k)$ and the conditional densities $P(x|\omega_k)$ are known. It is known that the joint probability density of finding a pattern that is in subclass ω_k and has feature value x can be written two ways: $P(\omega_k,x) = P(\omega_k|x)p(x) = P(x|\omega_k)P(\omega_k)$. Bayes' formula can be expressed as follows:

$$P(\omega_k \mid x) = \frac{P(x \mid \omega_k)P(\omega_k)}{P(x)} \tag{9}$$

Using (9), the decision rule of (8) can be rewritten as follows:

$$\frac{P(x \mid \omega_1)}{P(x \mid \omega_0)} \underset{\omega_0}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_0)}{P(\omega_1)} \tag{10}$$

The quantity on the left is called the likelihood ratio and is denoted by $\Lambda(x)$

$$\Lambda(x) \triangleq \frac{P(x \mid \omega_1)}{P(x \mid \omega_0)} \tag{11}$$

The quantity on the right-hand side of (10) is the threshold of the test and is denoted by η:

$$\eta \triangleq \frac{P(\omega_0)}{P(\omega_1)} \tag{12}$$

Thus, the Bayes criterion leads to the likelihood ratio test (LRT) shown in (13):

$$\Lambda(x) \underset{\omega_0}{\overset{\omega_1}{\gtrless}} \eta \tag{13}$$

Owing to the goodness of fit between the actual data and the theoretical data, the distributions of $P(x|\omega_0)$ and $P(x|\omega_1)$ are assumed to have an approximately exponential distribution:

$$P(x \mid \omega_k) = \alpha_k \times e^{-\beta_k x} \tag{14}$$

where k is 0 or 1 of each subclass ω, and α_k and β_k are the model's parameters. In this paper, the prior probabilities $P(\omega_0)$ and $P(\omega_1)$ for test sequences are investigated as shown in **Table 1**. On an average, $P(\omega_0)$ is 0.93, and $P(\omega_1)$ is 0.07. The model parameter values are α_0 = 1,140,000, β_0 = 2.824, α_1 = 2,810, and β_1 = 0.390.

Table 1. Prior probabilities according to test sequences

Genre	Seq.	$P(\omega_0)$	$P(\omega_1)$	Genre	Seq.	$P(\omega_0)$	$P(\omega_1)$
Lecture	A01	0.99	0.01	Religion &Documentary	B01	0.96	0.04
	A02	0.99	0.02		B02	0.93	0.07
	A03	0.99	0.01		B03	0.92	0.09
	A04	1	0		B04	0.86	0.14
	A05	1	0		B05	0.89	0.11

	A06	1	0.01		B06	0.92	0.08
	A07	0.99	0.01	Drama &	C01	0.81	0.19
	A08	0.98	0.02	Animation	C02	0.85	0.15
	A09	0.99	0.01		C03	0.91	0.09
	A10	0.94	0.06		C04	0.87	0.13
	A11	0.99	0.01	Music video & Sports	D01	0.91	0.09
	A12	0.93	0.07		D02	0.91	0.09
	A13	0.94	0.06		D03	0.82	0.18
	A14	0.97	0.03		D04	0.82	0.18
	A15	0.91	0.09		D05	0.88	0.12
				Average		0.93	0.07

Using (13), it can be determined whether the given two clusters are merged or not: two clusters are merged if $\Lambda(x)$ is greater than η. Finally, c clusters can be obtained according to FC_{intra} distribution, as shown in **Fig. 7**.

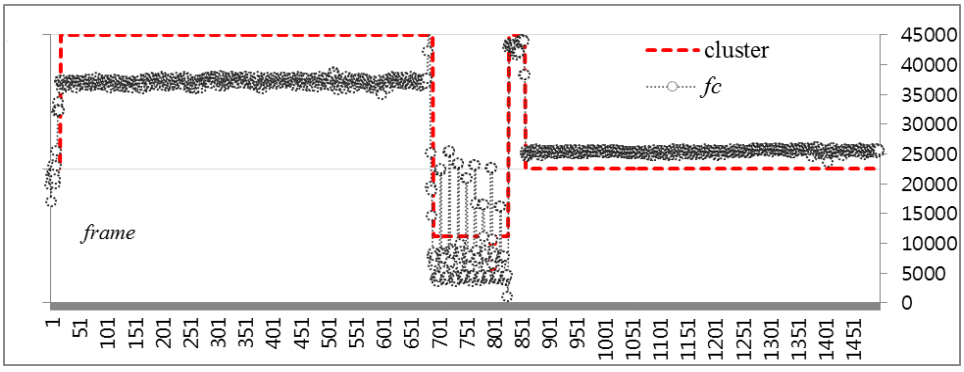


Fig. 7. Relationship between FC_{intra} distribution fc and the final clusters

Although the correlation between FC_{intra} and the number of bits is high, the maximum FC_{intra} frame does not always have the maximum number of encoded bits. Thus, the candidate intra-frame needs to be extracted. The candidate frame set $\mathbf{H_s}$ contains intra-frames, and a candidate frame in $\mathbf{H_s}$ is the frame that requires more than a certain number of encoded bits. $\mathbf{H_s}$ is specified in (15):

$$\mathbf{H}_{\mathrm{s}} = \{I_{q,1}^{c} \mid \theta(q) = i \text{ and } q = q+1, \text{ if } CC(Q_{s(i,1)}) \geq \mu_c \\ \text{where } 1 \leq q \leq D \leq M, \text{ and } 1 \leq i \leq M\}. \tag{15}$$

In (15), $I_{q,1}^{c}$ is a candidate intra-frame, D is the number of candidate frames, M is the number of intra-frames, $\theta(\bullet)$ is a nondecreasing mapping function from the integer set $\{1,\ldots,M\}$, and μ_c is the average of FC_{intra}'s in each cluster. If $CC(Q_{s(i,1)})$ is greater than the content-adaptive threshold μ_c, the ith intra-frame is extracted as $I_{q,1}^{c}$ of the cth cluster.

3.3 Model-Based Bit Rate Estimation

Using candidate frames with FC_{intra} value of each cluster, the bit rates of clusters can be estimated via statistical analysis and a mathematical model. To estimate the bit rate while maintaining the given PSNR quality, a PSNR-Q model derived from the H.264/AVC quantization process [15] is proposed in this paper. With this model, an estimated QP is determined and is finally applied to the bit rate estimation. The relationship between the quantization step size (*Qstep*) and QP is given in (16) as follows:

$$Qstep = \frac{2^{qbits} \times PF}{MF}, \tag{16}$$

where *PF* and *MF* are a post-scaling and a multiplication factor, respectively, in the H.264/AVC standard, and *qbits* = 15+floor (QP/6). When uniform quantization is applied to the uniformly distributed inputs, the mean square error (*MSE*) is given by

$$MSE = \frac{1}{Qstep}\int_{-Qstep/2}^{Qstep/2} u^2 du = \frac{Qstep^2}{12} \tag{17}$$

From (16) and (17), the PSNR can be derived as

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) = a \times QP + b \tag{18}$$

where *a* and *b* are constants obtained by linear regression [16]. As a result, the value of QP can be estimated as

$$QP_e = \frac{b - PSNR_t}{a} \tag{19}$$

where $PSNR_t$ is a given target PSNR, and QP_e is an estimated QP.

Using QP_e, the number of intra-frame bits is first estimated. Some parameters obtained by intra-frame estimation are used to estimate the number of inter-frames bits in a GOP. To estimate the number of intra-frame bits, a simple but effective Rate-Quantization (R-Q) model is used. An exponential relationship between the actual number of encoded bits and QP was modeled by Zhou and his colleagues [17]. For simplicity, the R-Q model for an intra-frame is defined as:

$$R_{q,1}(QP_e) = \alpha_q \times e^{(-\beta_q \times QP_e)} \tag{20}$$

where $R_{q,1}(QP_e)$ is the number of encoded bits for the *q*th candidate intra-frame at QP_e, and α_q and β_q are the model parameters. To reveal the relationship between the number of encoded bits and QP, **Fig. 8** shows several examples of curve-fitting results for intra-frames, with each small dot of the mathematically approximated curves representing the actual number of

encoded bits of an intra-frame at each QP. Because α_q and β_q can be obtained by exponential regression, $R_{q,1}$ can also be calculated by (20).

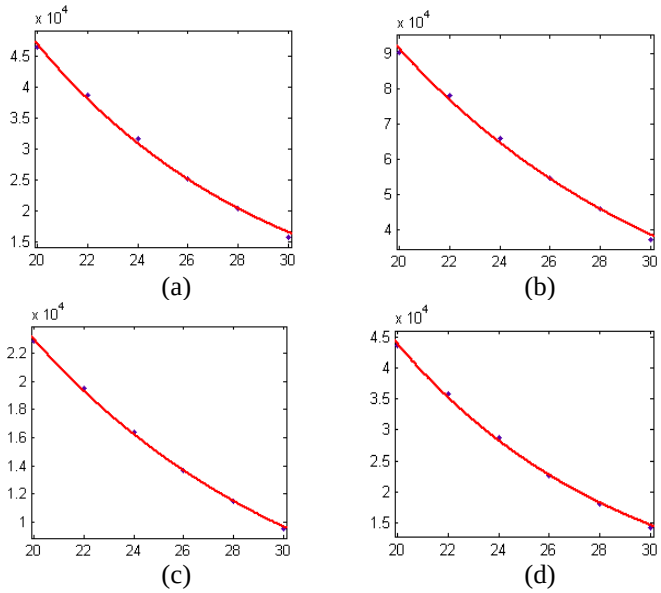


Fig. 8. R-Q curves for the test sequences. (a) Music video, (b) Lecture, (c) Sports, (d) Documentary

It is difficult to directly estimate the number of inter-frame bits in H.264/AVC. Thus, the bit rate conversion method introduced in [18] is used with the value of QP_e instead of using the intra-frame R-Q model. The bit rate conversion is defined as

$$R_{q,j+1}(QPP) = R_{q,j+1}(QP_s) \times 2^{(QPP-QP_s)/6}, \quad 1 \le j \le G-1 \tag{21}$$

where $R_{qj+1}(QPP)$ is the number of encoded bits for the (j+1)th inter-frame in the qth GOP at QPP, and G is a GOP size. As defined in (21), this method requires encoding a GOP at a certain value of QP, QP_s, as a reference, that is, $R_{q,j+1}$ (QP_s) is computed in advance. In experiments, the value of QP_s used is 26. Furthermore, QPP is set to QP_e+1 here because an inter-frame QP is an intra-frame QP+1 in H.264/AVC rate control. After estimating the number of intra- and inter-frame bits, the total number of bits for each GOP, R_q, can be estimated using (20) and (21) as follows:

$$R_q = R_{q,1} + \sum_{j=1}^{G-1} R_{q,j+1} \tag{22}$$

The bit rate of each cluster is estimated using the GOP that is expected to have the maximum number of encoded bits among all candidate frames in each cluster. If the same bit rate between clusters is estimated, these clusters are grouped as a segment. Finally, the number of segments in a sequence is less than or equal to the number of clusters.

4. Experimental Results

The performance of the proposed scheme is evaluated with several types of IPTV content. The proposed scheme will be called class-based bit rate estimation (CBRE) hereinafter, and the conventional scheme with a fixed bit rate of 2.5 Mbps will be called fixed bit rate estimation (FBRE) [19]. The standard definition (SD) resolution video content is categorized into four genres: lecture, religion and documentary, drama and animation, and music video and sports. A total of 30 videos in **Table 2** are used as test sequences.

Table 2. Test sequences

Class	Genre	Name	Run Time (m:s)	Name	Run Time (m:s)
Low-complexity video	Lecture	A01	(22:26)	A09	(27:07)
		A02	(23:54)	A10	(23:01)
		A03	(28:03)	A11	(26:46)
		A04	(27:51)	A12	(27:22)
		A05	(29:48)	A13	(25:54)
		A06	(30:54)	A14	(34:04)
		A07	(30:37)	A15	(35:40)
		A08	(23:41)	A16	(30:40)

Class	Genre	Name	Run Time (m:s)
High-complexity video	Religion & Documentary	B01	(26:15)
		B02	(53:38)
		B03	(52:33)
		B04	(25:18)
		B05	(56:51)
		B06	(14:58)
	Drama & Animation	C01	(23:08)
		C02	(80:24)
		C03	(16:33)
		C04	(14:07)
	Music Video & Sports	D01	(40:41)
		D02	(04:12)
		D03	(09:48)
		D04	(32:40)
		D05	(52:26)

In our experiment, the size of GOP is 15, and its type is set to IPPP. The target PSNR is set to 42dB. The simulated results encoded by FBRE can be compared in terms of the bit rate and quality to those encoded by CBRE. In order to evaluate the bit rate reduction, ΔR is calculated as follows:

$$\Delta R(\%) = \frac{1}{c}\sum_{i=1}^{c}\frac{R_i^{\mathrm{FBRE}} - R_i^{\mathrm{CBRE}}}{R_i^{\mathrm{FBRE}}} \times 100, \tag{23}$$

where R_i^{FBRE} and R_i^{CBRE} indicate the bit rates by FBRE and CBRE in the ith cluster, respectively.

Table 3 shows the results of bit rate reduction. CBRE can reduce the bit rate by up to 65.2% as compared with FBRE. CBRE can reduce the bit rate, on an average, by approximately 52% and 6% in low- and high-complexity video sequences, respectively. Because CBRE assigns the bit rate according to the complexity of each segment, a relatively high bit rate reduction in the low-complexity video class can be achieved.

Table 3. Bit rate reduction ratios of CBRE

Seq.	Segments	Estimated bit rates (kbps)	ΔR(%)
A01	4	743, 785, 1033, 1834	58.1
A02	6	382, 409, 504, 820, 1038, 1141	65.2
A03	5	336, 715, 1365, 1681, 2560	56.9
A04	4	401, 659, 1056, 1854	56.5
A05	5	251, 975, 986, 1595, 2560	51.6
A06	4	770, 995, 1120, 1559	59.5
A07	4	861, 1199, 1202, 1212	53.2
A08	4	310, 1042, 1175, 1472	54
A09	4	449, 577, 585, 1734	64.1
A10	4	507, 1015, 1115, 1283	55
A11	7	434, 541, 982, 1031, 1499, 1881, 2560	54.9
A12	4	637, 932, 1200, 1241	54.7
A13	4	1136, 1260, 1517, 2209	40.6
A14	3	500, 1705, 1770	32.9
A15	5	287, 1395, 1856, 2405, 2560	22.5
B01	4	343, 627, 848, 2560	1.3
B02	3	947, 2178, 2560	2.9
B03	3	1239, 1887, 2560	2.2
B04	2	1746, 2560	0.1
B05	1	2560	0
B06	1	2560	0
C01	3	980, 1405, 2560	1.7
C02	1	2560	0
C03	5	280, 1425, 1863, 2404, 2560	23.7
C04	1	2560	0
D01	4	237, 326, 1074, 2560	1.2
D02	2	1354, 2560	21.7
D03	5	322, 1396, 1893, 1896, 2560	26.4
D04	3	729, 1507, 2560	3.7
D05	2	412, 2560	0.1

Since the bit rate can be estimated by encoding candidate frames instead of the total frames, the computational complexity for CBRE depends on the ratio of the number of candidate frames to the total number of frames. **Fig. 9** shows these ratios in the test sequences.

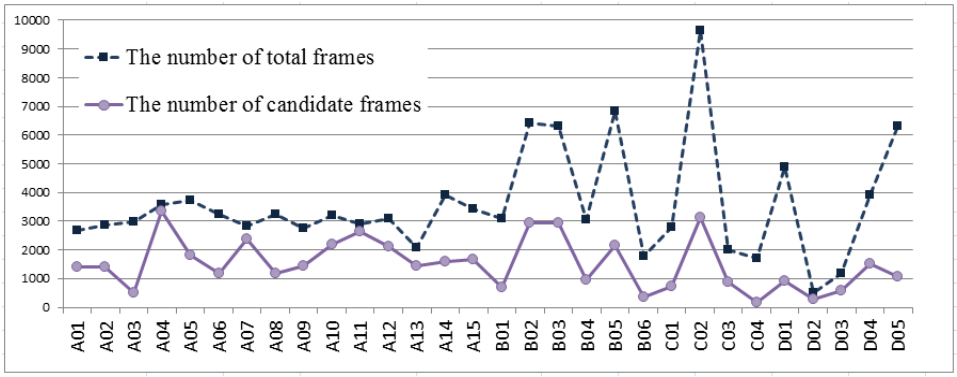


Fig. 9. Ratios of the number of candidate frames to the total number of frames in test sequences

Table 4 shows that the difference in the PSNR performance is approximately 1.2dB on an average. However, that is too small a difference to affect the subject quality degradation in test sequences as shown in **Fig. 10**, since the target bit rate is set to 40dB in (19), which makes it difficult to determine a subjective quality difference.

Table 4. PSNR difference between FBRE and CBRE

Seq.	PSNR(dB)		Seq.	PSNR(dB)	
	R_i^{FBRE}	R_i^{CBRE}		R_i^{FBRE}	R_i^{CBRE}
A01	45.97	43.5	B01	42.39	42.3
A02	46.24	43.1	B02	43.54	43.5
A03	46.95	44.5	B03	43.62	43.6
A04	47.16	45.1	B04	40.2	40.2
A05	46.71	44.2	B05	41.84	41.84
A06	49.64	43.1	B06	41.35	41.35
A07	47.13	44.4	C01	46.05	46
A08	46.57	44.5	C02	39.93	39.93
A09	46.87	44.2	C03	44.8	44.3
A10	46.37	44.3	C04	42.23	42.23
A11	46.2	44.3	D01	42.71	42.7
A12	46.1	45.2	D02	45.6	44.9
A13	45.65	44.3	D03	45.39	44.7
A14	43.99	42.8	D04	43.97	43.7
A15	46.5	45.9	D05	44.32	44.3
Average PSNR Drop			-1.2		

(a)

(b)

Fig. 10. Subjective quality comparison: (a) CBRE and (b) FBRE

5. Conclusions

The transcoding bit-rate decision in conventional IPTV and Smart TV services is a time-consuming process since the input sequence should be fully transcoded several times with different bit-rates to decide a suitable bit-rate. This paper shows that the video bit rate in an MPEG-2 TS to H.264/AVC transcoder which is an essential device in those services can be automatically decided with keeping subjective video quality. The proposed bit rate estimation scheme was organized into two modules: one was hierarchical clustering-based sub-classification and the other was statistical analysis-based bit rate estimation. The input sequence was grouped as several subclasses by hierarchical clustering using the parameter value extracted from each frame. The candidate frames of each subclass were used to estimate the bit rate using statistical analysis and mathematical model. The bit rate could be automatically estimated by encoding only the candidate frames.

The proposed scheme could reduce the fixed bit rate, on an average, by 52% in low-complexity video and by 6% in high-complexity video while maintaining the subjective quality, respectively. For future work, we plan to study some practical issues for implementing the proposed scheme. Note that in real TV services, additional works need to be developed in order to simplify the proposed scheme, especially clustering-based video sub-classification. We also need to extend the results to HD test sequences.

References

[1] H. L. Cycon, T. C. Schmidt, M. Wahlisch, D. Marpe, and M. Winken, “A temporally scalable video codec and its applications to a video conferencing system with dynamic network adaption for mobiles,” *IEEE Trans. Consumer Electron.*, vol. 57, no. 3, pp. 1408-1415, Aug. 2011. Article (CrossRef Link).

[2] S. Park, and S. H. Jeong, “Mobile IPTV: approaches, challenges, standards and QoS support,” *IEEE Internet Comput.*, vol. 13, no. 3, pp. 22-31, May–Jun. 2008. Article (CrossRef Link).

[3] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560-576, Jul. 2003. Article (CrossRef Link).

[4] A. Joch, F. Kossentini, H. Schwarz, T. Wiegand, and G.J. Sullivan, “Performance comparison of video coding standards using Lagrangian coder control,” *in Proc. of IEEE Int. Conf. Image Processing*, vol. 2, pp. II-501-504, Sep. 2002. Article (CrossRef Link).

[5] T. Kim and H. Bahn, “Implementation of the storage manager for an IPTV set-top box,” *IEEE Trans. Consumer Electron.*, vol. 54, no. 4, pp. 1770-1775, Nov. 2008. Article (CrossRef Link).

[6] ITU-R Recommendation BT.500-11, “Methodology for the subjective assessment of the quality of

television pictures," ITU, 2002. Article (CrossRef Link).
[7] Wook Joong Kim, Jong Won Yi, and Seong Dae Kim, "A bit allocation method based on picture activity for still image coding," *IEEE Trans. Image Process.*, vol. 8, no. 7, pp. 974-977, 1999. Article (CrossRef Link).
[8] J. Li and E. Abdel-Raheem, "Efficient rate control H.264/AVC intra frame," *IEEE Trans. Consumer Electron.*, vol. 56, no. 5, pp. 1043-1048, May 2010. Article (CrossRef Link).
[9] X. Jing, L.-P. Chau, and W.-C. Siu, "Frame complexity-based rate-quantization model for H.264/AVC intraframe rate control," *IEEE Trans. Signal Process. Lett.*, vol. 15, pp. 373-376, 2008. Article (CrossRef Link).
[10] Y. Zhou, Y. Sun, Z. Feng, and S. Sun, "New rate-distortion modeling and efficient rate control for H.264/AVC video coding," *Signal Process.: Image Commun.*, vol. 24, no. 5, pp. 345-356, May 2009. Article (CrossRef Link).
[11] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, 2nd ed., Wiley-Interscience, 2000, pp. 20-82. Article (CrossRef Link).
[12] C.Y. Chiu, C.S. Chen, and L.F. Chien, "A framework for handling spatiotemporal variations in video copy detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 3, pp. 412-417, Mar. 2008. Article (CrossRef Link).
[13] G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, pp. 91-110, 2004. Article (CrossRef Link).
[14] M. M. Deza and E. Deza, *Encyclopedia of Distances*, 1st ed., Springer, 2009, pp. 89-100. Article (CrossRef Link).
[15] Y. Liu, Z. G. Li, and Y. C. Soh, "A novel rate control scheme for low delay video communication of H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 68-78, Jan. 2007. Article (CrossRef Link).
[16] A.L. Edwards, *An Introduction to Linear Regression and Correlation*, W.H. Freeman, pp. 33-46, 1976. Article (CrossRef Link).
[17] Y. Zhou, Y. Sun. Z. Feng, and S. Sun, "New rate-distortion modeling and efficient rate control for H.264/AVC video coding," *Signal Process.: Image Commun.*, vol. 24, no. 5, pp. 345-356, May 2009. Article (CrossRef Link).
[18] Q. Tang, H. Mansour, P. Nasiopoulos, and R. Ward, "Bit-rate estimation for bit-rate reduction H.264/AVC video transcoding in wireless networks," *in Proc. of IEEE Int. Sym. Wireless Pervasive Comput.*, pp. 464-467, May 2008. Article (CrossRef Link).
[19] H. J. Cho, J. Lee, D. Y. Noh, S. H. Jang, J. C. Kwon, and S. J. Oh, "A new video bit rate estimation scheme using a model for IPTV services," *KSII Trans. Internet and Information Syst.*, vol. 5, no. 10, pp. 1814-1829, Oct. 2011. Article (CrossRef Link).

Hye Jeong Cho received the B.S. degree in 2004 from the Department of Internet Information Engineering, Hanyang Women's College, Seoul, Korea. In 2012, she received the joint M.S. and Ph.D. degree in electronic engineering, Kwangwoon University, Seoul, Korea. She is currently a senior engineer in AV Research and Development Laboratory, ARION Technology Inc., Gyeonggi-do, Korea. Her research interests include video processing, STB and IPTV video streaming services.

Chae-Bong Sohn received the B.S., M.S., and Ph.D. degree in electronic engineering from Kwangwoon University, Seoul, Korea in 1993, 1995, and 2006, respectively. He is currently an associate professor in department of Electronics and Communications Engineering, Kwangwoon University, Seoul, Korea. His research interests include image compression, transcoding, digital broadcasting systems.

Seoung-Jun Oh was born Seoul, Korea, in 1957. He received both the B.S. and the M.S. degrees in electronic engineering from Seoul National University, Seoul, in 1980 and 1982, respectively, and the Ph.D. degree in electrical and computer engineering from Syracuse University, New York, in 1988. In 1988, he joined ETRI, Daejeon, Korea, as a senior research member. From 1990 to 1992, he was a Director of Multimedia Research Section, ETRI. Since 1992, he has been a professor of Department of Electronic Engineering, Kwangwoon University, Seoul, Korea. He has been a chairman of SC29-Korea since 2001. His research interests include image and video processing, video coding, and object recognition.