JOINT CONTROL FOR HYBRID TRANSCODING USING MULTIDIMENSIONAL RATE DISTORTION MODELING

Yong Ju Jung and Yong Man Ro

Multimedia Group, Information and Communications University (ICU)
Yusong, Daejeon, PoBox. 77, 305-732, Korea

ABSTRACT

In this paper, we present the hybrid transcoding, specifically, practical joint control of spatial, temporal, and SNR scaling by rate-distortion modeling. The objective of the joint control is to determine the optimal combination of transcoding operations. To allocate bits more efficiently, the multidimensional factors including spatial, SNR, and temporal and their tradeoffs are exploited in terms of the quality of content. Based on the statistical analysis of dependency distortion, we propose an improved distortion model and a rate control algorithm. In experiments, a typical video transcoding problem is analyzed and solved by using the proposed rate-distortion (R-D) model.

1. INTRODUCTION

Content adaptation is an important technique for maximizing content accessibility, e.g., it could provide pervasive and interactive functionalities. Since various networks might have different bandwidths, a gateway could include a transcoder or modality converter adapting the video bit rates in order to provide consistent video services to users. Specifically, content transcoding allows the multimedia content to be adapted to a wide diversity of client device capabilities in communication, processing, and display as well. Users could have the best quality possible. Typically, in a decision for content adaptation, one should take into account information about multimedia content, network characteristics, device capabilities and user preferences. Final objective is to provide the best presentation to user given a set of constraints related with information in above.

The motivation of our research is to find an optimal transcoding strategy for the best perception with the given content and resource constraints, *i.e.* to find a possible set of transcoding operators to meet the bit rate constrained by network or terminal for the best quality of the adapted content. Furthermore, in order to allocate bits more efficiently, the multidimensional (spatio-SNR-temporal) tradeoffs should be exploited. For example, in the case of reducing bitrate, it should be decided to transmit either more contents with lower SNR quality or less contents with higher quality [3] or either more contents with smaller picture size or less contents with larger one. Various tradeoffs should be considered to jointly control the spatial, temporal, and SNR scaling.

The paper is organized as follows: In Section 2, we mention briefly about video transcoding and formulize the problem. In Section 3, we discuss a distortion modeling and a rate control algorithm based on the R-D modeling. In Section 4, we show experimental results and the effectiveness of the proposed rate control. Finally, we conclude the paper in Section 5.

2. HYBRID VIDEO TRANSCODING

Most previous works focused on either bit allocation over coded frames under a constant frame rate or frame rate control at a fixed spatial resolution. Those could not provide the transcoding decision considering fully spatio-temporal-SNR trade-offs. In previous work, the trade-offs between SNR and temporal quality were studied, where the trade-offs were achieved with a simple parametric model [1]. In [2], a multi-dimensional bit rate control for selecting video coding parameter was studied. However, they did not focus on model based rate control but operational one. Also, in [3], a joint control technique for SNR-temporal factors was proposed by adding one more constraint to the conventional rate-quantizer (R-Q) model. In [4], authors attempted to model the distortion for multidimensional transcoding and coding. However, they did not consider the dependency among transcoding operations and the systematic joint rate control for hybrid transcoding.

In this paper, we focus on the multidimensional transcoding, specifically, practical joint control of spatial, temporal, and SNR scaling by rate-distortion modeling.

2.1. Problem formulation

The R-D optimization problem for the hybrid transcoding having three dimensional quality control points in the above is to find an optimal operation set, {**RQ***,**TS***,**SS***}, so that the average distortion of the transcoded content is minimized.

$$\{\mathbf{RQ}^*, \mathbf{TS}^*, \mathbf{SS}^*\} = \arg \min_{\mathbf{RQ},\mathbf{TS},\mathbf{SS}} \sum_{i=1}^{N} D_i(\mathbf{RQ}, \mathbf{TS}, \mathbf{SS}),$$

$$subject\ to\ \sum_{i=1}^{N} R_i(\mathbf{RQ}, \mathbf{TS}, \mathbf{SS}) < R_{\max} \quad (1)$$

where D_i is the distortion of the i^{th} frame, which is represented by mean square error (MSE). In (1), $\mathbf{RQ}$ represents a requantization vector, i.e., $\mathbf{RQ} = [(RQ_1, RQ_2, \cdots, RQ_i) \mid i = 1,...,N]$, $RQ_i \in [RQ_{\min}, RQ_{\max}]$, where RQ_i is the requantization parameter, which can be any value of the quantization step between $RQ_{\min}$ and $RQ_{\max}$, and N is the total number of frames in a temporal segment (e.g., scene); $\mathbf{TS}$=$[(TS_1, TS_2, \cdots, TS_i) \mid i = 1,..., N], TS_i \in \{0,1\}$, where $\mathbf{TS}$ is a temporal scaling operation vector, and TS_i is an integer value to represent temporal dropping, where 1 means no-dropping and 0 does frame dropping; and $\mathbf{SS} = [(SS_1, SS_2, \cdots, SS_i) \mid i = 1,..., N], SS_i \in (0,1]$, where $\mathbf{SS}$ is a spatial resolution scaling operation vector, and $SS_i = \dfrac{n_1' \times n_2'}{n_1 \times n_2}$, where n_1 x n_2 is the spatial resolution of the original coded frame, and $n_1^{'}$ x $n_2^{'}$ is the downscaled spatial resolution.

To solve this optimization problem, the R-D modeling can provide practical solutions that are able to generate automatic adaptation policies with reasonable performance. In what follows, the distortion for the hybrid video transcoding is modeled, and a joint rate control algorithm based on the model is introduced.

3. JOINT RATE CONTROL FOR HYBRID TRANSCODING

3.1. Rate-distortion model

In conventional R-D modeling, there are various models obtained by considering the statistical characteristics of source. For Laplacian source, $p(x)=\frac{\alpha}{2}e^{-\alpha|x|}$ with $-\infty<x<\infty$, a quadratic R-D model was adopted in MPEG-4 [8]. In this paper, we also assume that the source model is Laplacian model. Then, the distortion (MSE (mean square error)) due to (re)quantization operation for a frame can be written as

$$E\{(\frac{1}{\alpha e^{R}})^2\}, where\, R>0 \tag{2}$$

where R is bits/pixel.

To simplify the analysis of the distortion due to the temporal dropping, we assume a simple repeat interpolation (zero-order hold) to estimate the dropped frame. Then, the distortion (MSE) generated by the temporal interpolation can be modeled with the spatial gradients and motion vectors (mv_x, mv_y) from the previous coded frame, which is obtained from the optical flow equation [6], i.e.,

$$E\{(\frac{\partial \tilde{s}}{\partial x}mv_x+\frac{\partial \tilde{s}}{\partial y}mv_y)^2\}\,. \tag{3}$$

The distortion (MSE) generated by the spatial interpolation can be calculated as follows:

$$E\{(\tilde{s}(x,y)-\bar{s}(x,y))^2\} \tag{4}$$

where $\tilde{s}\,(x,y)$ is a coded pixel at (x,y) and $\bar{s}(x,y)$ is a pixel value interpolated from the (sub)sampled value at (x,y). On the same condition of simple repeat interpolation, this distortion is simply modeled with the local variance of source.

In general, the distortion of the i^{th} frame can be represented as the weighted sum of the distortions which are generated by each transcoding operation.

$$D_i(\mathbf{RQ},\mathbf{TS},\mathbf{SS})=w_{rq}D_i^{rq}(\mathbf{RQ},\mathbf{TS},\mathbf{SS})+w_{ts}D_i^{ts}(\mathbf{RQ},\mathbf{TS},\mathbf{SS}) +w_{ss}D_i^{ss}(\mathbf{RQ},\mathbf{TS},\mathbf{SS}) \tag{5}$$

where D_i^{rq}, D_i^{ts}, and D_i^{ss} refer to the distortion of requantization, temporal scaling, and spatial resolution scaling operations, respectively, and the weighting factors $\{w_{rq},w_{ts},w_{ss}\}\in[0,1]$, and $w_{rq}+w_{ts}+w_{ss}=1$.

3.1.1. Independent distortion model

In case that there is no inter-frame and inter-operation dependency, $D_i^{rq}(\mathbf{RQ},\mathbf{TS},\mathbf{SS})$, $D_i^{ts}(\mathbf{RQ},\mathbf{TS},\mathbf{SS})$, and $D_i^{ss}(\mathbf{RQ},\mathbf{TS},\mathbf{SS})$ depend only on any single variable of RQ_i, TS_i, and SS_i. Therefore, $D_i(\mathbf{RQ},\mathbf{TS},\mathbf{SS})$ can simply be the weighted sum of three distortions arising from RQ_i, TS_i, and SS_i as in [7], i.e.,

$$D_i(\mathbf{RQ},\mathbf{TS},\mathbf{SS})=w_{rq}D_i^{rq}(RQ_i)+w_{ts}D_i^{ts}(TS_i)+w_{ss}D_i^{ss}(SS_i)\,. \tag{6}$$

To simplify (6), one can consider two cases, the distortion at the coded frame and the distortion at the dropped frame. As in [5] and [6], the distortion is defined as the weighted sum of coded frame distortion and dropped frame distortion, i.e.,

$$D_i(\mathbf{RQ},\mathbf{TS},\mathbf{SS}) =\begin{cases} w_{ts}D_i^{ts}(TS_i), & for\, dropped\, frame(TS_i=0) \\ w_{rq}D_i^{rq}(RQ_i)+w_{ss}D_i^{ss}(SS_i), & for\, coded\, frame(TS_i=1). \end{cases} \tag{7}$$

If we assume that a temporal interpolator repeats the previous transcoded frame $\hat{s}_p$, then the estimated error is

$$e_i=\tilde{s}_i-\hat{s}_i=\tilde{s}_i-\hat{s}_p=(\tilde{s}_i-\tilde{s}_p)+(\tilde{s}_p-\hat{s}_p) \tag{8}$$

where $\hat{s}_i$ is the transcoded frame at the i^{th} frame, and $\hat{s}_i=\hat{s}_p$ by a simple interpolator, and $\tilde{s}_i-\tilde{s}_p$ represents the frame interpolation error, and $\tilde{s}_p-\hat{s}_p$ is the transcoding error of the previous transcoded frame. Then, the distortion due to the temporal scaling operation can be rewritten as follows:

$$D_i^{ts}(TS_i)=E\{e_i^2\}\,. \tag{9}$$

With the assumption of independence, (9) becomes

$$E\{(\tilde{s}_i-\tilde{s}_p)^2+(\tilde{s}_p-\hat{s}_p)^2\}\,. \tag{10}$$

Obviously, since $D_p^{ts}(TS_p)=0$ for coded frame, (10) becomes $E\{(\tilde{s}_i-\tilde{s}_p)^2\}+w_{rq}D_p^{rq}(RQ_p)+w_{ss}D_p^{ss}(SS_p)$. (11)

Therefore, the distortion of the i^{th} frame, i.e., (7) can be rewritten as follows:

$$D_i(\mathbf{RQ},\mathbf{TS},\mathbf{SS}) =\begin{cases} w_{ts}E\{(\frac{\partial \tilde{s}_i}{\partial x}mv_{x_{p,i}}+\frac{\partial \tilde{s}_i}{\partial y}mv_{y_{p,i}})^2\}+w_{ts}w_{rq}E\{(\frac{1}{\alpha_p e^{R_p}})^2\} \\ \quad +w_{ts}w_{ss}E\{(\tilde{s}_p(x,y)-\bar{s}_p(x,y))^2\},\ TS_i=0 \\ w_{rq}E\{(\frac{1}{\alpha_i e^{R_i}})^2\}+w_{ss}E\{(\tilde{s}_i(x,y)-\bar{s}_i(x,y))^2\},\ TS_i=1. \end{cases} \tag{12}$$

3.1.2. Dependent distortion model

In the inter-operation dependent case, one transcoding operation can affect the distortion caused by the other transcoding operations. To model this inter-operation dependent distortion, we address first an analysis of statistical properties of DCT coefficients. Figure 1 shows AC coefficients distribution of coded frames and spatially transcoded ones. As seen from Fig. 1, the spatial downsampling operation affects the source distribution, i.e., α is changed for the Laplacian source statistics, i.e., $p(x)=\frac{\alpha'}{2}e^{-\alpha'|x|}$. The rate-distortion function in (2) is correspondingly changed as α is changed. That means the R-D function of requantization operation for SNR scaling is affected by the amount of spatial resolution downsampling. Moreover, we observed that a frame dropping operation also affects the source distribution like the temporal dependency in motion compensated video coding. The amount of impact relies on the GOP structure and the previous dropped reference frames. Based on the above observations, we can conclude as follows: due to the dependency among transcoding operations, temporal and spatial resolution scaling operations affect the traditional R-D model of SNR. Therefore, the dependency distortion should be compensated by considering the amount of change in the parameter of R-D function for requantization operation.

Let $D_i^{rq(\alpha_i')}$ be the (re)quantization distortion of the i^{th} frame with new alpha value α_i' of the source distribution model which is changed due to the spatial resolution scaling and/or temporal scaling operation. The distortion due to requantization operation with inter-operation dependency can be defined as follows:

$$D_i^{rq}(\mathbf{RQ},\mathbf{TS},\mathbf{SS})=\begin{cases}0, & TS_i=0\\ D_i^{rq(\alpha_i')}(RQ_i), & TS_i=1.\end{cases} \quad (13)$$

There is no inter-operation dependency between temporal and spatial resolution scaling operations. Further, the spatial distortion in the i^{th} frame is not related to the distortion in previous frames. Therefore, the distortion for the spatial resolution scaling operation can be written as

$$D_i^{ss}(\mathbf{RQ},\mathbf{TS},\mathbf{SS})=\begin{cases}0, & TS_i=0\\ D(SS_i), & TS_i=1.\end{cases} \quad (14)$$

As in the independent case, the distortion due to the temporal scaling operation is the sum of the frame interpolation distortion and the transcoding distortion of the previous frame, i.e.,

$$D_i^{ts}(\mathbf{RQ},\mathbf{TS},\mathbf{SS})=E\{(\tilde{s}_i-\tilde{s}_p)^2\}+D_p(\mathbf{RQ},\mathbf{TS},\mathbf{SS}). \quad (15)$$

$$=\begin{cases}E\{(\frac{\partial\tilde{s}_i}{\partial x}mv_{x_{p,i}}+\frac{\partial\tilde{s}_i}{\partial y}mv_{y_{p,i}})^2\}+w_{rq}E\{(\frac{1}{\alpha_p'e^{R_p'}})^2\}\\ \quad +w_{ss}E\{(\tilde{s}_p(x,y)-\bar{s}_p(x,y))^2\}, TS_i=0\\ 0, \quad TS_i=1.\end{cases} \quad (16)$$

Now, by using (13), (14), and (16), the distortion for the i^{th} frame, i.e., (5) is given by

$$D_i(\mathbf{RQ},\mathbf{TS},\mathbf{SS})$$
$$=\begin{cases}w_{ts}E\{(\frac{\partial\tilde{s}_i}{\partial x}mv_{x_{p,i}}+\frac{\partial\tilde{s}_i}{\partial y}mv_{y_{p,i}})^2\}+w_{ts}w_{rq}E\{(\frac{1}{\alpha_p'e^{R_p'}})^2\}\\ \quad +w_{ts}w_{ss}E\{(\tilde{s}_p(x,y)-\bar{s}_p(x,y))^2\}, \quad TS_i=0\\ w_{rq}E\{(\frac{1}{\alpha_i'e^{R_i'}})^2\}+w_{ss}E\{(\tilde{s}_i(x,y)-\bar{s}_i(x,y))^2\}, TS_i=1.\end{cases} \quad (17)$$

3.2. Rate control for hybrid transcoding

Figure 2 shows the proposed segment-level rate control algorithm. Using this rate control algorithm, we can get the optimal combination of transcoding operations consisting of the frame rate, the scaling factor of resizing, and the amount of average requantization for a segment by finding an operation set that minimizes the average distortion. We assume that an input video is segmented at first by using some conventional segmentation method. In Step 1 of the algorithm, we calculate the bitrate by using the following bitrate estimation for each operation. To estimate the rate due to a requantization operation, the following conventional quadratic rate-quantization model is used [8]: $R_i(RQ_i)=\frac{a}{RQ_i^2}+\frac{b}{RQ_i}$ (18)

where a and b are the model parameters. To estimate the rate due to a temporal scaling operation, we need to obtain the number of bits for each frame in the segment. When a frame is dropped, if we know the number of bits for the frame, we can easily calculate the bitrate due to the temporal scaling operation. To estimate the rate due to a spatial resolution scaling operation, we can use an intuitive rule. When a spatial resolution is downscaled by two in both horizontal and vertical sizes, i.e., TS_i=1/4, the bitrate is empirically reduced by 3 to 4. The reduction ratio strongly depends on the downscaling algorithm and video sequence.

In MPEG case, since the amount of distortion due to the requantization operation depends on frame type, each average amount of requantization for an I frame type, a P frame type, and a B frame type in a segment should be obtained. With regard to computational complexity, the number of operations is generally finite. Therefore, the computational complexity is not high. Regarding the computation of the model parameters to estimate the rate and distortion, since we do not deal with encoding but transcoding, the model parameters are easily obtained during the transcoding process. Parameters α', a, and b can be obtained by regression analysis [8]. The motion information, i.e., mv_x, mv_y, the local variance of source and the number of bits for each frame can be directly extracted or quickly estimated from bitstream.

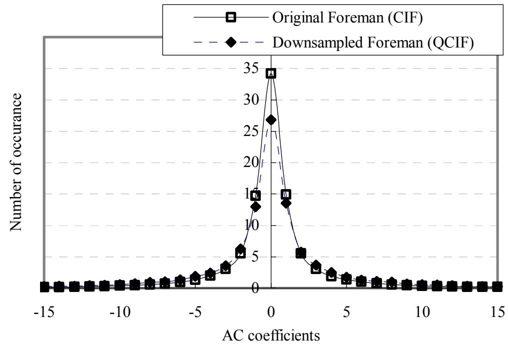

Fig. 1. Probability mass function of AC coefficients. Frame 0 of Foreman CIF originally coded at 30fps and 510kbps.

Step 0: Calculate the target bit rate for a segment;
Step 1: Estimate each bitrate for all the possible combinations of transcoding operations which consist of the frame rate, the scaling factor of resizing, and the amount of average requantization for the segment;
Step 2: Select the probable operation sets from all the operation combinations in Step 1, which can generate the target rate;
Step 3: Estimate the total distortion for the selected operation sets in Step 2 by using the distortion measure (17);
Step 4: Select an operation set which minimizes the total distortion by solving (1);

Fig. 2. Proposed rate control algorithm for a segment.

4. EXPERIMENTAL RESULTS

In our experiments, we have used Foreman, Container Ship, and Football video sequences that are originally encoded as MPEG-

4 advanced simple profile format with 30f/s and CIF resolution. The sequences have I-B-P GOP structure (M=3, N=15). We assume that the spatial downsampling ratio is the same for each frame in a temporal video segment, i.e., $SS_1 = SS_2 = \cdots = SS_N = S_c$, where $0 < S_c \le 1$. We also assume that frame dropping is uniform, and then we define dropping parameter f_s to represent the amount of the frame dropping. That is, $f_s = \frac{F_S}{F_T}$, where F_S is the original frame rate of the source video, and F_T is the frame rate of the transcoded video. In our experiments, the transcoding operations used for hybrid transcoding are as follows: average requantization parameter $\overline{RQ} \in \{1,..,31\}$, spatial resolution downsampling parameter $S_c \in \{1, 1/4\}$, and temporal resolution subsampling parameter $f_s \in \{1, 2/3, 1/2, 1/3, 1/5\}$. This means that we transcode the bitstream as the type of {CIF, QCIF} and {30f/s, 20f/s, 15f/s, 10f/s, 5f/s} with various SNR qualities by requantization.

Table 1 shows the possible operations which can generate the target rate 80Kbps from CIF Foreman originally coded at 510Kbps. In our experiment, we have used the scene unit as the segment for the segment level rate control. Foreman sequence consists of 3 scenes. This result is for the first scene which is from frame 1 to frame 170. As seen from Table 1, the best operation set which minimizes the distortion is (18, 1, 1/4). To verify the effectiveness of our joint control, we examine the actual rate of the transcoded bitstream, the estimated distortion, and the finally selected operation pair satisfying each target rate. Also, we examine the actually measured mean square error. Table 2 shows the result of that for hybrid transcoding. This result shows that the distortion is estimated well.

Table 3 and 4 show the possible operations for transcoding to the target bitrate from Container Ship and Football each. Container Ship has low motion, so neighboring frames are very similar. This means that the temporal downsampling causes the small amount of distortion. Therefore, the best operation set (29, 1/5, 1) is a reasonable choice. On the contrary, Football has high motion. This means that the little amount of temporal downsampling is better. Our result (14, 1, 1/4) matches well this estimation.

5. CONCLUSION

In this paper, we have proposed the practical joint control of multidimensional transcoding by the rate-distortion model based approach. Also, we have discussed tradeoffs in quality among transcoding operations to allocate bits more efficiently. By describing the statistical analysis of dependency distortion, we have proposed an improved distortion model.

REFERENCES

[1] F.C. Martins, W. Ding, and E. Feig, "Joint control of spatial quantization and temporal sampling for very low bit rate video," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, pp. 2072-2075, May 1996.

[2] E.C. Reed, and J.S. Lim, "Optimal Multidimensional Bit-Rate Control for Video Communication," *IEEE Trans. Image Processing*, vol. 11, no. 8, pp. 873-885, Aug. 2002.

[3] A. Vetro, H. Sun, and Y. Wang, "Object-Based Transcoding for Adaptable Video Content Delivery," *IEEE, Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 387-401, Mar. 2001.

[4] P. Yin, A. Vetro, M. Xia, and B. Liu, "Rate-Distortion Models for Video Transcoding," in *Proc. SPIE-IS&T Electronic Imaging,* vol. 5022, pp. 479-488, Jan. 2003.

[5] S. Liu, and C.-C. J. Kuo, "Joint Temporal-Spatial Rate Control for Adaptive Video Transcoding," in *Proc. Int. Conf. Multimedia and Expo*, vol. 2, July 2003.

[6] J.-W. Lee, A. Vetro, Y. Wang, and Y.-S. Ho, "Bit Allocation for MPEG-4 Video Coding With Spatio-Temporal Tradeoffs," *IEEE, Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 488-502, June 2003.

[7] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit Allocation for Dependent Quantization with Applications to Multiresolution and MPEG Video Coders," *IEEE Trans. Image Processing*, vol.3, pp. 553-545, Sept. 1994.

[8] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate-distortion modeling," *IEEE, Trans. Circuits Syst. Video Technol.*, Feb. 1997.

Table 1. Estimated operations which can generate the target rate 80 kb/s from Foreman originally coded at 510 kb/s with CIF and 30f/s

$\overline{RQ}$	f_s	S_c	Estimated MSE
7	1/5	1/4	496.37
10	1/3	1/4	442.37
12	1/2	1/4	368.00
14	2/3	1/4	292.98
18	1	1/4	141.11
21	1/5	1	445.91
30	1/3	1	387.07

Table 2. Results of joint control for hybrid transcoding of Foreman originally coded at 510 kb/s with CIF and 30f/s

Target rate (Kb/s)	Actual transcoded rate (Kb/s)	Estimated MSE	Selected operation set ($\overline{RQ}, f_S, S_C$)	Actual MSE
30	31	519.05	(25, 1/5, 1/4)	396.94
40	43	380.71	(30, 1/2, 1/4)	235.38
60	73	146.39	(28, 1, 1/4)	193.52
80	87	141.11	(18, 1, 1/4)	155.24
100	103	137.22	(13, 1, 1/4)	140.58
200	278	72.54	(22, 1, 1)	76.20
300	324	68.10	(15, 1, 1)	67.26
400	420	63.49	(11, 1, 1)	49.08
500	454	59.99	(9, 1, 1)	47.29

Table 3. Estimated operations which can generate the rate 50 kb/s from Container Ship originally coded at 262 kb/s with CIF and 30f/s

$\overline{RQ}$	f_s	S_c	Estimated MSE
12	1/5	1/4	286.58
15	1/3	1/4	294.11
17	1/2	1/4	292.18
19	2/3	1/4	289.03
22	1	1/4	277.09
29	1/5	1	165.50

Table 4. Estimated operations which can generate the target rate 300 kb/s from Football originally coded at 946 kb/s with CIF and 30f/s

$\overline{RQ}$	f_s	S_c	Estimated MSE
5	1/5	1/4	2920.88
7	1/3	1/4	2468.30
9	1/2	1/4	1909.77
11	2/3	1/4	1914.95
14	1	1/4	175.05
14	1/2	1	2857.05
21	1/3	1	2410.32
27	1/2	1	1838.40