Subjective Assessment for Resolution Improvement on 4K TVs

- Analysis of Learning-Based Super-Resolution and Non-Linear Signal Processing Techniques -

Hiroki Shoji† Seiichi Gohshi‡
†‡Department of Information Science
Kogakuin University
Tokyo, Japan
e-mail: †em15011@ns.kogakuin.ac.jp, ‡gohshi@cc.kogakuin.ac.jp

Abstract—**Super-resolution (SR) is a technology to create high-definition images. According to television (TV) manufacturer's advertisements, TVs sold recently in Japan have SR functions. In Japan, when such TVs are sold, SR is aggressively advertised on a large scale; however, in countries other than Japan, SR is not mentioned in similar TV manufacturer's advertisements. In previous research, real-time processing to generate SR images has been found to be difficult. It is necessary to verify whether SR advertised by TV manufacturers exhibits its original performance in a TV that requires a real-time processing. However, an objective assessment of SR on TVs cannot be conducted because images processed in TVs cannot be extracted. Therefore, in our previous work, a subjective assessment of Learning-Based Super-Resolution (LBSR) was conducted, and it was shown that the subjective assessment is effective in performance verification of SR on a TV. Moreover, we conducted a subjective assessment of LBSR and Non-Linear Signal Processing (NLSP) using a 4K TV to evaluate the image quality of each SR image produced via up-conversion from HD video to 4K. In this study, the image quality of each SR when improving resolution of 4K video is evaluated by the subjective assessment. Furthermore, the performance results of both LBSR and NLSP for resolution improvement on a 4K TV are reported.**

Keywords—Learning-Based Super-Resolution; Non-Linear Signal Processing; 4K TV; Subjective Assessment; Performance Verification.

I. INTRODUCTION

Super-Resolution (SR) is a technology for improving the resolution of images and videos. In recent years, research and development of SR for 4K television (TV) has been increasingly active. Most 4K TVs currently sold have SR functions; SR on a 4K TV is used to up-convert low-resolution content to 4K. SR on 4K TV is used to up-convert low-resolution content to 4K. Broadcasting and Blu-ray content typically use high-definition (HD) resolution and because 4K content has been insufficient until only a few years ago, most content for 4K TV must be up-converted from HDTV content. However, recently, 4K content is increasing and is being streamed over the Internet. Further, test broadcasting for the practical use of 4K broadcasting has been actively conducted. Therefore, we expect 4K content to be increasingly common in the future.

SR can also improve resolution. When 4K content becomes more widespread in the future, SR on 4K TV will be needed for resolution improvement. SR is uniquely developed by TV manufacturers to include resolution improvement functions. Most TV manufacturers focus their development of SR on Learning-Based Super-Resolution (LBSR) [1][2][3].

In Japan, when TVs are sold, SR is aggressively advertised on a large scale. However, in countries other than Japan, SR is not mentioned in similar TV manufacturer's advertisements [4][5][6][7]. It is necessary to verify whether SR advertised by TV manufacturers exhibits the original performance in the TV that requires real-time processing.

Performance of SR is generally measured using Peak Signal-to-Noise Ratio (PSNR). The authors conducted an objective assessment using PSNR about performance of LBSR [8]. In [8], real-time processing of SR has been found to be difficult. However, the performance of SR developed by the TV manufacturers is not published in advertisements. Additionally, the objective assessment of SR on a TV cannot be conducted because images processed in the TV cannot be extracted. Therefore, the authors conducted a subjective assessment to measure performance of SR on a TV [9]. In [9], performance of LBSR and Non-Linear Signal Processing (NLSP) [10] were evaluated when HD video was up-converted to 4K. It is possible to compare performance of each SR by analyzing statistically subjective assessment data. Accordingly, the subjective assessment is effective in performance verification of SR on a TV. In related research, subjective assessments of SR image Reconstruction (SRR) and NLSP have been completed using methods that up-convert (i.e., HD to 4K) and resolution improvement (i.e., 4K to 4K) [11][12]. Further, a subjective assessment of LBSR in up-converting HD to 4K was completed and compared to NLSP; however, LBSR performance for resolution improvement is yet to be evaluated.

Therefore, in this study, we focus on a subjective assessment of resolution improvement. The subjective assessment comprises an experiment for collecting data subjectively assessed by study subjects. Assessment targets of our experiment comprise the following three methods: a 4K original signal; NLSP; and LBSR. The collected assessment data is statistically analyzed and LBSR performance on a 4K TV is quantitatively shown. Significance tests using Analysis of Variance (ANOVA) and a yardstick graph are then performed; a significant difference between each the technique is obtained. Finally, we prove that LBSR is inferior to NLSP and conclude that NLSP is useful as a SR resolution improvement technique for 4K TV.

The paper is structured as follows. In Section II, the subjective assessment experiment is explained. In Section III, experiment results are analyzed by statistical methods are explained. In Section IV, the analyzed results are discussed. Finally, Section V presents a conclusion about this study.

II. Subjective Assessment Experiment

In this section, the subjective assessment method and the experiment overview are explained.

A. *Subjective assessment method*

In [9][11][12], subjective assessments via paired comparisons are conducted. In this study, to quantitatively assess LBSR performance for resolution improvement, Scheffe's paired comparison, ANOVA, and the yardstick graph are adopted. Each of these methods is described in the next section.

B. *Experimental method*

Assessment targets in this study are OFF (i.e., the original 4K signal), NLSP, and LBSR. The subjective assessment method is Scheffe's paired comparison in which assessment pairs are created and compared to one another when three or more assessment targets are present. A relative comparison in this experiment is a method to assess the other target using a five-step scale (i.e., -2 to 2) when one side of the targets is a criterion (i.e., 0 points). The assessment scale is shown in Table 1, with the five steps defined as Excellent, Good, Fair, Poor, and Bad. As an example, when evaluating NLSP as compared to LBSR, an assessment score of 2 is assigned if NLSP has a higher definition than LBSR, 0 if NLSP is the same as LBSR, and -2 if NLSP has a lower definition than LBSR. Here, high definition is a state in which fine components of a given video are more clearly displayed.

Assessment data obtained via the subjective assessment are then applied to a significance test using ANOVA. Further, experimental results having significant differences are ranked via the yardstick graph.

In the subjective assessment of the video, there is a possibility that the assessment score is changed because of the evaluation order. Therefore, it is conducted our experiments using various combinations to increase the reliability of the assessment data. More specifically, one subject assesses the following six patterns: NLSP and LBSR when the criterion is OFF; OFF and LBSR when the criterion is NLSP; and OFF and NLSP when the criterion is LBSR. Because the subjects are not experts, an oral description regarding the resolution of the video is provided before each experiment, and the subjects understand the differences in resolution via a demonstration. Moreover, in this experiment, because the subjects provide their assessments by replaced the criterion, they often get confused. Therefore, the subjects are instructed to assess only after correctly understanding the given criterion target. Further, the subjects are instructed to ignore the differences in color temperature, color tone, and noise during the reproduction.

TABLE 1. SUBJECTIVE ASSESSMENT SCALE

Assessment score	Assessment word	Description of assessment words (as compared to a reference)
2	Excellent	Very good resolution
1	Good	Good resolution
0	Fair	Degree resolution is the same
-1	Poor	Bad resolution
-2	Bad	Very bad resolution

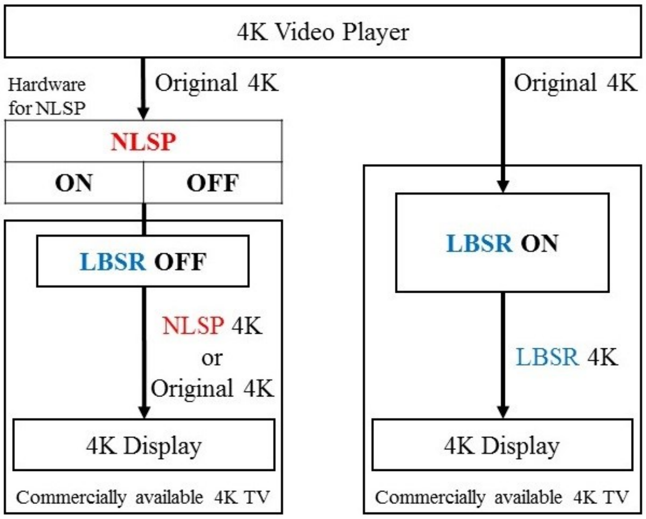


Figure 1. Block diagram of our experimental equipment

Figure 2. NLSP hardware

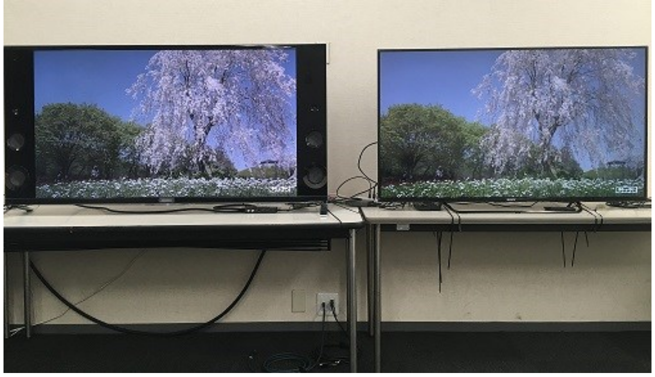


Figure 3. Assessment targets
(Left TV is OFF or NLSP, Right TV is LBSR)

ISBN: 978-1-61208-513-5

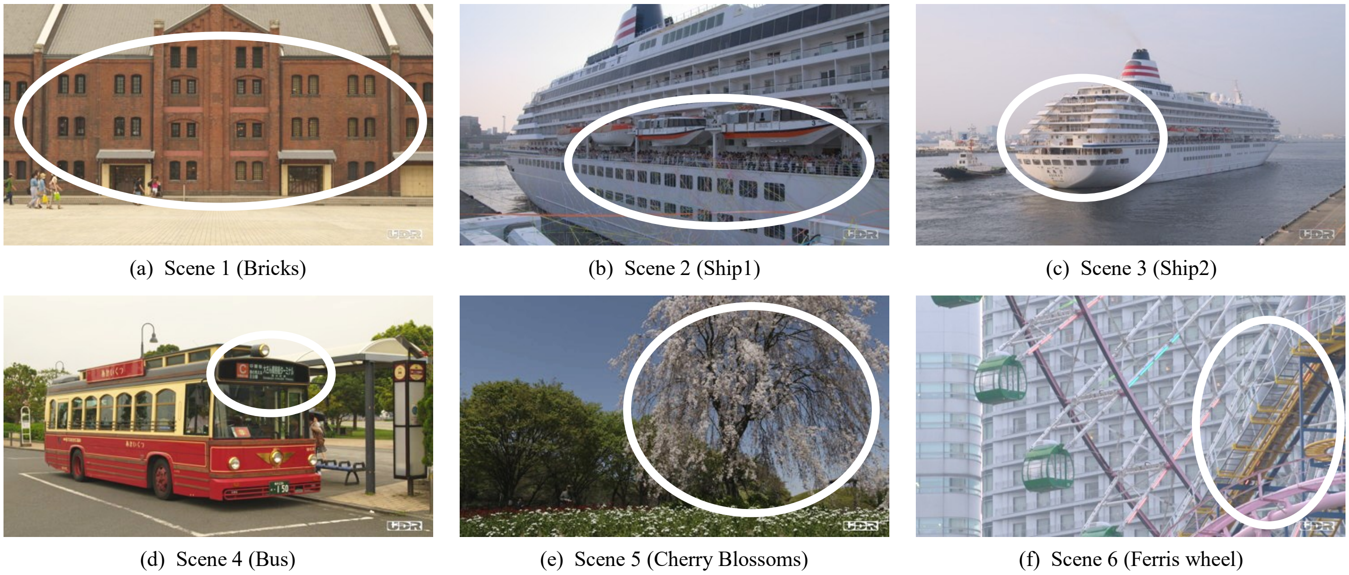

(a) Scene 1 (Bricks)
(b) Scene 2 (Ship1)
(c) Scene 3 (Ship2)
(d) Scene 4 (Bus)
(e) Scene 5 (Cherry Blossoms)
(f) Scene 6 (Ferris wheel)

Figure 4. Experimental 4K videos

C. Experimental equipment

In this section, our experimental equipment is described. Figure 1 shows a block diagram summarizing our experimental equipment. In the figure, the HDTV player is able to reproduce video in an uncompressed form, unlike a conventional DVD player. Although experimental videos were recorded in MPEG-4 format, such videos were never compressed during reproduction while using this player. Details of the experimental videos are described later.

Figure 2 shows the NLSP hardware; here, NLSP and OFF are able to switch a single TV ON and OFF via this hardware. An indicator displaying ON-OFF is present, enabling us to understand whether the subjects have watched either NLSP results or the 4K original signal. Here, ON indicates NLSP, whereas OFF indicates the 4K original signal.

As shown in Figure 3, two 4K TVs are used in this experiment. Here, the manufacturers of the two 4K TV sets are the same, but because the model numbers differ, each 4K TV's color temperature and color tone differs slightly. The liquid crystal panel does not exist exactly the same thing, even if the model number or the product lot are the same. Therefore, using different model numbers is not problem.

D. Experimental videos

The experimental videos were shot using a consumer 4K video camera with fine components to easily confirm differences in resolution but also with coding deterioration of MPEG-4. Here, flickers or deformations of high-frequency components are caused by the coding degradation. In this experiment, videos that included these degradations are assessed. Note that there were no large movements such as panning or tilting in any of the experimental videos. Reproduction time was 10-15 seconds and each video is looped. The input resolution was 4K resolution (i.e., 3840 × 2160) and was improved to 4K resolution by each of the resolution enhancement processes. Figure 4 summarizes the videos used in our experiments. Regions indicated by white circles in the figure include a fine pattern of bricks in Scene 1, passengers and details of window frames in Scene 2, the appearance and character of a ship in Scene 3, the characters on a bus in Scene 4, fineness of petals in Scene 5, and fineness of the framework in Scene 6. All such scenes help the subjects to easily confirm the differences in resolution. The subjects performed their assessments while primarily watching these regions.

E. Experimental subjects

Experimental subjects are 30 non-experts, both men and women of 20s with no problems in visual acuity, color vision, and field of view.

F. Experimental environment

As shown in Figure 5, to reproduce the environment in which a consumer selects a TV in a shop, the viewing environment is bright. Although the viewing distance was not fixed, the subjects always assessed the TVs by standing in front of them.

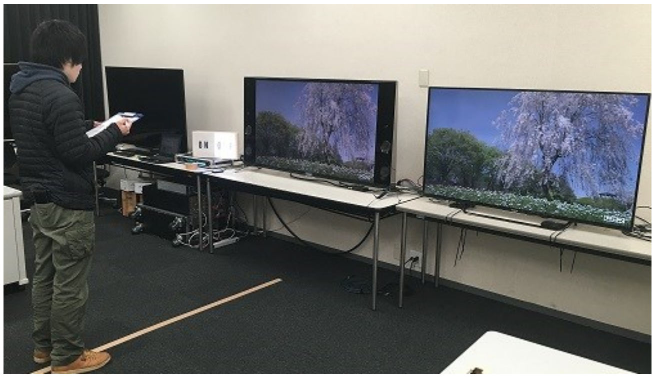


Figure 5. Experimental environment

ISBN: 978-1-61208-513-5

TABLE 2. CROSS TABLES

(a) Cross Table of Scene 1 (Bricks)

	OFF	NLSP	LBSR	Xi
OFF		56	18	74
NLSP	-54		-22	-76
LBSR	-22	30		8
Xj	-76	86	-4	X…
Xj-Xi	-150	162	-12	6

(b) Cross Table of Scene 2 (Ship1)

	OFF	NLSP	LBSR	Xi
OFF		55	15	70
NLSP	-56		-23	-79
LBSR	-9	33		24
Xj	-65	88	-8	X…
Xj-Xi	-135	167	-32	15

(c) Cross Table of Scene 3 (Ship2)

	OFF	NLSP	LBSR	Xi
OFF		49	11	60
NLSP	-47		-21	-68
LBSR	-12	22		10
Xj	-59	71	-10	X…
Xj-Xi	-119	139	-20	2

III. ANALYSIS AND RESULTS

In this study, the assessment data obtained in the subjective assessment are analyzed and statistically quantified. Below, the results of our analysis is presented.

A. Cross table

A cross table is used to organize the assessment data. The cross tables shown in Table 2 provide summed values of the assessment data. In Table 2, typical results are shown and other results are similar to these results. Below, the use of a cross table is explained using Scene 1 of Table 2. In the table, OFF, NLSP, and LBSR in the first column show the criterion methods, whereas OFF, NLSP, and LBSR in the first row show the assessment targets. Each value is the sum of the assessment scores of each subject. For example, the score of 56 in row two, column three is the sum of the assessment scores for NLSP when the criterion is OFF. Conversely, the score of -54 in row three, column two is the sum of the assessment scores of OFF when the criterion is NLSP. In general, Xi is the sum of the assessment scores in each row and Xj is the sum of the assessment scores in each column. Further, X… is the sum of Xi and Xj. As an example, -150, 162, and -12 scores in the sixth row (i.e., Xj - Xi) are calculated from the difference of each Xj and Xi; these values are used for ANOVA and the yardstick graph.

B. Analysis of variance (ANOVA)

Table 3 shows our typical results of ANOVA. Using Scene 1 of Table 3, the ANOVA table is explained. In this study, the factors analyzed by ANOVA are the main effect, main effect × individual, combination, order effect, and order effect × individual; these are shown in rows two through six of the ANOVA table. The factor shown represents the cause that affected each assessment score. The seventh row is a residual, and the eighth row is a total. The second column is the sum of squares (S), the third column is the degree of freedom (DoF), and the fourth column is variance (V). The main effect is calculated as follows:

$$S = \frac{1}{2nN}\sum(Xj - Xi)^2 \qquad (1)$$

$$DoF = n - 1 \qquad (2)$$

$$V = S/DoF \qquad (3)$$

In (1) and (2), n is the number of assessment targets and N is the number of subjects. In our experiments, n = 3 and N = 30 are set. Further, the values shown in Table 2 as Xj - Xi values are used. In the fifth column, F represents the variance ratio, which is the quotient obtained by dividing the variance of each factor by the residual. As an example, F (506.41) of the main effect in Scene 1 of Table 2 was calculated by dividing variance (135.80) by residual (0.27); however, an error occurs if F is calculated using this value, because the values in Table 3 are rounded off. In the sixth column, F1% represents the variance ratio (i.e., boundary value) of each factor with a significance level of 1% calculated using the FINV function of Excel. For the significance test of ANOVA, we used the F value of the main effect, noting a significant difference at F > F1%. As an example, in Scene 1 of Table 3, F was 506.41 and F1% was 4.85. Here, because F is larger than F1%, a significant difference exists between the assessment targets with a significance level of 1%. Similar to Scene 1, in Scenes 2 through 6, because F is larger than F1%, the presence of a significant difference has successfully been shown.

TABLE 3. ANALYSIS OF VARIANCE (ANOVA) TABLES.

※DoF: Degree of freedom

(a) ANOVA Table of Scene 1 (Bricks)

Factor	Sum of squares	DoF	Variance	F	F1%
Main	271.60	2	135.80	506.41	4.85
Main × Individual	40.40	58	0.70	2.60	1.60
Combination	1.80	1	1.80	6.71	6.93
Order	0.20	1	0.20	0.75	6.93
Order × Individual	6.13	29	0.21	0.79	1.93
Residual	23.87	89	0.27	-	-
Total	344.00	180	1.91	-	-

(b) ANOVA Table of Scene 2 (Ship1)

Factor	Sum of squares	DoF	Variance	F	F1%
Main	261.88	2	130.94	439.85	4.85
Main × Individual	41.79	58	0.72	2.42	1.60
Combination	5.34	1	5.34	17.93	6.93
Order	1.25	1	1.25	4.20	6.93
Order × Individual	4.25	29	0.15	0.49	1.93
Residual	26.49	89	0.30	-	-
Total	341.00	180	1.89	-	-

(c) ANOVA Table of Scene 3 (Ship2)

Factor	Sum of squares	DoF	Variance	F	F1%
Main	188.23	2	94.12	251.29	4.85
Main × Individual	57.43	58	0.99	2.64	1.60
Combination	5.00	1	5.00	13.35	6.93
Order	0.02	1	0.02	0.06	6.93
Order × Individual	7.98	29	0.28	0.73	1.93
Residual	33.33	89	0.37	-	-
Total	292.00	180	1.62	-	-

TABLE 4. SCALE VALUE TABLES

(a) Scale value of Scene 1 (Bricks)

Target	OFF	NLSP	LBSR
Scale value (α)	-0.83	0.90	-0.07

(b) Scale value of Scene 2 (Ship1)

Target	OFF	NLSP	LBSR
Scale value (α)	-0.75	0.93	-0.18

(c) Scale value of Scene 3 (Ship2)

Target	OFF	NLSP	LBSR
Scale value (α)	-0.66	0.77	-0.11

(d) Scale value of Scene 4 (Bus)

Target	OFF	NLSP	LBSR
Scale value (α)	-0.76	0.89	-0.13

(e) Scale value of Scene 5 (Cherry Blossoms)

Target	OFF	NLSP	LBSR
Scale value (α)	-0.85	1.01	-0.16

(f) Scale value of Scene 6 (Ferris wheel)

Target	OFF	NLSP	LBSR
Scale value (α)	-0.76	0.83	-0.07

C. *Yardstick*

Given that a significant difference is proved in the main effect via ANOVA, detailed significance tests are conducted between each assessment target via the yardstick graph. Accordingly, existence of significant difference is proved visually.

Here, a scale value is calculated to create the yardstick graph. The scale value in this study quantifies the performance of the resolution enhancement processing of the assessment target, wherein the height of this value represents the height of performance. The scale value is calculated as follows:

$$\alpha = \frac{1}{2\mathrm{nN}}(Xj - Xi) \qquad (4)$$

Here, n, N, and Xj - Xi are the same as above. Table 4 shows the scale values for each experimental video.

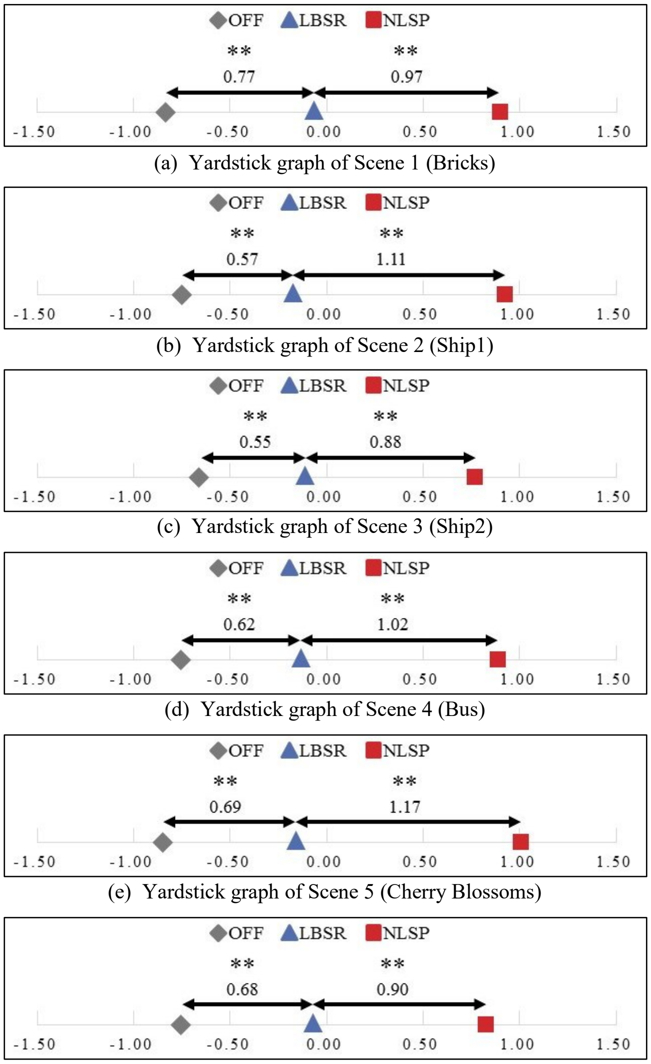
A graph using this scale as its horizontal axis is a yardstick graph. Figure 6 shows yardstick graphs for each experimental video. In the figure, a rhombus indicates OFF, a triangle indicates LBSR, and a square indicates NLSP. Values between the assessment targets represent the distances of the scale value. According to Figure 6, for all experimental videos, scale value ranking was OFF, LBSR, and NLSP in ascending order. Regarding the distance between the assessment targets, differences between NLSP and the other two methods were large, which indicated a particularly good performance of NLSP in resolution enhancement processing. Similarly, it can be confirmed that LBSR was better than OFF.

From the yardstick graph, performance differences are quantitatively showed. To confirm the presence of significant differences in these performance differences, a significance test is conducted using assessment standard value Ya, as calculated below.

$$Y_{\alpha} = q\sqrt{\frac{V_{\varepsilon}}{2nN}} \qquad (5)$$

Here, q is the q value of the studentized range, Vε is the variance of the residual shown in the ANOVA table, and n and N are the same as above. Table 5 shows assessment standard values Y1% for all experimental videos with significance levels of 1%.

A significant difference is observed in significance level 1% when the distance between the assessment targets was greater than Y1%. In Scene 1 of Figure 6, the distance between NLSP and LBSR (0.97) is bigger than Y1% (0.16). In addition, the distance between LBSR and OFF (0.77) is also



(a) Yardstick graph of Scene 1 (Bricks)

(b) Yardstick graph of Scene 2 (Ship1)

(c) Yardstick graph of Scene 3 (Ship2)

(d) Yardstick graph of Scene 4 (Bus)

(e) Yardstick graph of Scene 5 (Cherry Blossoms)

(f) Yardstick graph of Scene 6 (Ferris wheel)

Figure 6. Yardstick graphs

TABLE 5. ASSESSMENT STANDARD VALUES Y1%

	Y1%
Scene 1	0.16
Scene 3	0.19
Scene 5	0.15

	Y1%
Scene 2	0.17
Scene 4	0.16
Scene 6	0.17

bigger than Y1% (0.16). Therefore, the significant difference is observed in significance level 1%. The two asterisks in each yardstick graph of Figure 6 represent the existence of a significant difference with a significance level of 1%. As a result of the significance tests for all experimental videos, a significant difference is found with a significance level of 1% between NLSP and LBSR, as well as LBSR and OFF. This result shows that resolution enhancement processing of NLSP and LBSR is statistically effective.

IV. Discussion

Based on the results obtained from our experiments and analysis, it is discussed about the significance of each resolution enhancement method.

LBSR has significant differences in the significance level of 1% as compared with that of the 4K original signal, thus showing the resolution enhancement processing of LBSR to be statistically effective. Further, according to our analysis results of the yardstick graphs, the performance of NLSP is better than LBSR. More specifically, we statistically and quantitatively showed that NLSP is better than LBSR.

In a recent study, using deep convolutional neural networks, more advanced LBSR techniques have been proposed [13] on the premise of applying such techniques to still images. As long as the approach is learning based, processes will require longer processing times, such as for the analysis of an input image, a database search, and block matching. Therefore, LBSR does not meet the real-time requirements for TV. In LBSR on a TV, manufacturers expect that a dedicated large-scale integrated processor could solve the problem of real-time processing; however, there is a limit to what can be solved via hardware. To realize effective real-time processing, it can be considered that there is a possibility that some process has been simplified. On the other hand, NLSP that we have proposed is able to create components are exceeded the Nyquist frequency in real-time. It has been proved in [10]. In addition, NLSP is able to process in real-time even if it is mounted on a conventional simple device because it is very simple signal processing. Therefore, it can be said that a hardware cost is low.

In addition, from the opinions provided by our test subjects, problems in NLSP and LBSR need to be solved. In the videos processed by LBSR, image artifacts are present. Such artifacts are image disturbances such as block noise and aliasing. When conducting subjective assessments, it was necessary to select areas that did not have many artifacts as one's focal point. Therefore, in LBSR, we require processing to reduce aliasing and other such artifacts. In the experimental video of the cherry blossoms with many high-frequency components processed by NLSP, we heard opinions noting that a subject's eyes were tired because of excessive emphasis on the image. Therefore, we conclude it necessary to find optimum processing parameters for each video.

In the future, after resolving the aforementioned problems, we plan to conduct further subjective assessments for various types of images. We also plan to increase the accuracy of our evaluation experiment. In particular, we conclude that videos with a face and text are preferred.

V. Conclusions

In this study, we conducted a subjective assessment of NLSP and LBSR on 4K TV. Analyzing the assessment results obtained in our experiments, we quantitatively showed the performance of NLSP and LBSR incorporated into a 4K TV. It was found that NLSP was better than LBSR. Further, it was found that LBSR was statistically more effective as compared with the original 4K signals. Therefore, we conclude that SR for resolution improvement on 4K TVs is indeed effective. In the future, we plan to implement more accurate assessment experiments by increasing the number and variety of assessment videos.

References

[1] http://www.sony.jp/bravia/featured/picture.html [retrieved: Oct, 2016]

[2] http://panasonic.jp/viera/technology/hexa_chroma/remaster.html [retrieved: Oct, 2016]

[3] http://www.lg.com/jp/lgtv/4k-ultrahdtv#colum2 [retrieved: Oct, 2016]

[4] http://www.sony.com/electronics/4k-resolution-4k-upscaling-tvs [retrieved: Oct, 2016]

[5] http://shop.panasonic.com/tvs/4k-tvs [retrieved: Oct, 2016]

[6] http://www.lg.com/us/tvs [retrieved: Oct, 2016]

[7] http://www.samsung.com/us/video/tvs [retrieved: Oct, 2016]

[8] H. Shoji, and S. Gohshi, "Limitations of Learning-Based Super-Resolution," 2015 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2015), pp.646-651, Nov.2015.

[9] H. Shoji, and S. Gohshi, "Performance of Learning-Based Super-Resolution on 4K-TV," The 47th ISCIE International Symposium on Stochastic Systems Theory and Its Applications (SSS'15), pp.79-80, Dec.2015.

[10] S. Gohshi, "Realtime Super Resolution for 4K/8K with Nonlinear Signal Processing," Journal of SMPTE (Society of Motion Pictures and Television Engineers), 124, pp. 51-56, Oct. 2015.

[11] M. Sugie, S. Gohshi, H. Takeshita, and C. Mori, "Subjective Assessment of Super-Resolution 4K Video using Paired Comparison," 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2014), pp.42-47, Dec.2014.

[12] C. Mori, M. Sugie, H. Takeshita, and S. Gohshi, "Subjective Assessment of Super-Resolution: High-Resolution Effect of Nonlinear Signal Processing," 10th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT 2015), IEICE & IEEE, pp.46-48, Aug.2015.

[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.295-307, June.2015.