PERFORMANCE EVALUATION OF IMAGE QUALITY METRICS WITH RESPECT TO THEIR USE FOR SUPER-RESOLUTION ENHANCEMENT

Tomáš Lukeš, Karel Fliegel, Miloš Klíma
Department of Radioelectronics, Faculty of Electrical Engineering
Czech Technical University in Prague, Czech Republic

ABSTRACT

Super-resolution (SR) methods enable to obtain high resolution image (HR) from multiple low resolution input images (LR) of the same scene. Development of SR algorithms for real applications require reliable image quality metrics for performance comparison of various SR methods. In this paper, six standard SR methods were implemented and subjective image quality assessment was performed to evaluate the visual quality of the enhanced images. Nine image quality metrics were examined with respect to their ability to characterize the observed subjective image quality of SR enhancement.

Index Terms—*Super-resolution, image processing, image quality metrics, subjective image quality assessment*

1. INTRODUCTION

Objective image quality metrics can be divided into two categories: full-reference metrics (FR) and no-reference metrics (NR). In real applications, the reference image is unknown. Therefore, no-reference metrics for SR should be preferred. Several studies have been made to evaluate performance of SR methods [1], [2], [3]. In this paper, the task is to determine the most suitable image quality metric for SR reconstruction. Based on the subjective assessment, nine image quality metrics were compared. The ideal objective metric for SR should be no-reference, computationally inexpensive, easily accessible, easy to use and in close agreement with human judgments.

2. SUPER-RESOLUTION METHODS USED

A review of super-resolution methods can be found in [4]. Four implemented algorithms are based on non-uniform interpolation. Shift and add algorithm (ShiftAdd) is a variation of the nearest neighbor approach. The bilinear SR (bilinSR) method uses the bilinear weighted sum to calculate each HR grid point. Delaunay triangulation [5] is a core of the third implemented algorithm (DelTr). The near optimal (near opt.) non-uniform interpolation [6] derives the weights for the closest pixels using synthetic LR images generated from an arbitrary HR image. Next algorithm is based on iterative back projection (IBP) [7] and the MAP [8] algorithm estimates the final HR image by maximizing the probability that it was formed from the input LR images.

3. SUBJECTIVE QUALITY ASSESSMENT

Since in real applications there is no ground truth that could be used as a reference, single stimulus (SS) adjectival categorical judgment method has been employed with respect to ITU-R BT.500 recommendation [9]. Five levels (excellent, good, fair, poor, bad) ITU-R quality scale has been used. Only one subject per session was assessing the test images on a calibrated monitor (EIZO CG242W). Seven different images[1] were processed by six implemented SR algorithms and bilinear interpolation to create a set of test images. Thirty-six non-expert observers participated in the test (all without visual impairment). The age ranged from 20 to 25. At the beginning of the session the test methodology and the range of quality levels was explained through a set of training examples. Test images were zoomed to 200%. It allows more precise examination of image artifacts. It also well simulates the common situation when the viewer zooms on a detail of the image in an image browser.

Using the outlier removal procedure described in [9], one of the subjects has been discarded as an outlier. Figure 1 shows the mean opinion score (MOS) for all test images. The 95% confidence interval according to the ITU-R recommendation is given by $[MOS_j - \sigma_j,\ MOS_j + \sigma_j]$, $\sigma_j = 1.96\, S_j/\sqrt{N}$, where N is the number of observers and S_j is the standard deviation of the test condition j across the observers.

4. OBJECTIVE QUALITY METRICS COMPARISON

Six FR and three NR metrics were calculated (Table 1) and compared by Spearman rank order correlation coefficient (SROCC), Pearson's linear correlation coefficient (PLCC) and root mean square error (RMSE) between MOS and objective metric score after nonlinear regression. Four parameter logistic function which is a part of IVQUEST evaluation software [10] was used for the regression.

This work was partially supported by the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services - QUALINET, by the COST CZ LD12018 Modeling and verification of methods for Quality of Experience (QoE) assessment in multimedia systems - MOVERIQ and by the grant No. P102/10/1320 Research and modeling of advanced methods of image quality evaluation of the Czech Science Foundation.

[1] Some input images are from http://r0k.us/graphics/kodak/. Database of test images used in this paper together with MOS ratings is freely available for research purposes at http://dbq.multimediatech.cz/ctusr/

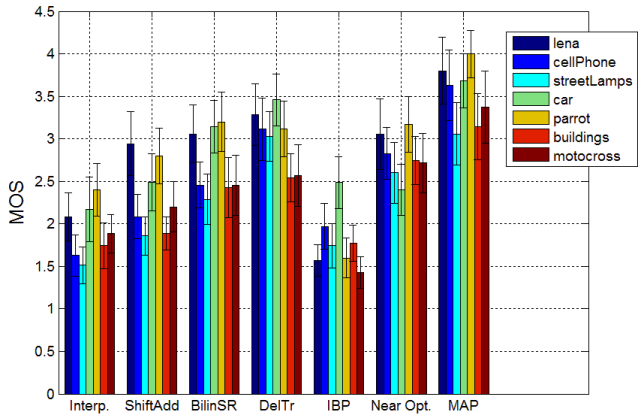

Fig. 1: Results of the subjective assessment: MOS and confidence intervals vs. standard interpolation (interp.) and 6 SR methods.

Three state of the art NR metrics theoretically applicable to SR images were examined. Metric Q provides very good results for parameter optimization of image denoising algorithm [16], but it seems not suitable to compare SR methods. Details are often perceived due to high frequency image content, so sharpness may be a clue about SR image quality. Therefore, CPBD [17] image sharpness metric was tested. The imperfect implementation of IBP [7] method causes undesirable artifacts like sharp lines across the edges. It confuses the sharpness based metrics and deteriorates CPBD's [17] performance. NIQE [15] provides promising results. However, its performance varies depending on the input parameters (32 x 32 px block size has provided the best results in this case).

Table 1: Objective metrics comparison with respect to SR images

	PLCC	**SROCC**	**RMSE**	**Type**
SSIM [11]	0.663	0.650	0.492	FR
MS-SSIM [11]	0.782	0.743	0.409	FR
PSNR	0.595	0.594	0.528	FR
VIFP [12]	0.706	0.696	0.465	FR
VSNR [13]	0.609	0.615	0.521	FR
UIQI [14]	0.606	0.576	0.522	FR
NIQE [15]	0.585	0.647	0.532	NR
Metric Q [16]	0.319	0.492	0.622	NR
CPBD [17]	0.402	0.528	0.601	NR

5. CONCLUSIONS

Subjective quality assessment has been done by evaluating 49 images created by six SR methods and bilinear interpolation. Nine image quality metrics were compared to determine the most suitable metric for evaluation of SR methods. Imperfections in SR reconstruction cause undesirable artifacts like ringing, additional noise and most importantly pixelation across the edges. All tested metrics have achieved rather low correlation with MOS, probably because they are not well tuned to consider SR artifacts that appear signicantly different than common image compression artifacts. Pixelation across the edges strongly deteriorates subjective image quality. The majority of tested metrics is not able to evaluate correctly this change in the image structure. MS-SSIM full reference metric offers the best performance, because it takes into account structural distortions. NIQE has reached the best results among NR metrics and approximates to FR.

In our future research we would like to extend the test image database and perform additional subjective assessments. Next research effort could be devoted to the development of a new, unified, no-reference image quality metric specialized for SR. The key factor for further improvement of current metrics is to evaluate better the SR artifacts as pixelation across the edges.

6. REFERENCES

[1] K. Nelson. A. Bhatti. S. Nahavandi. "Performance Evaluation of Multi-frame Super-resolution Algorithms,"DICTA 2012 International Conference, Dec. 2012.

[2] A. R. Reibman. R. M. Bell. S.Gray. "Quality assessment for super-resolution image enhancement," IEEE International Conference on Image Processing. ICIP, pp. 2017-2020, Oct. 2006.

[3] A. R. Reibman. T. Schaper. "Subjective performance evaluation of super-resolution image enhancement," Second Int. Workshop on Video Proc. and Qual. Metrics (VPQM'06), Jan. 2006.

[4] S. C. Park. M.K. Park. M.G. Kang. "Super-Resolution Image Reconstruction: A Technical Overview". IEEE Signal Processing Magazine, vol. 20. pp. 21-36, May 2003.

[5] S. Lertrattanapanich. N. K. Bose. "High resolution image formation from low resolution frames using Delaunay triangulation". IEEE Trans. Image Process., vol. 11. pp. 1427 – 1441, Dec. 2002.

[6] A. Gilman. D. G. Bailey. "Near optimal non-uniform interpolation for image super-resolution from multiple images," Massey University, Palmerston North, 2006.

[7] M. Irani. S. Peleg. "Improving resolution by image registration," CVGIP: Graph Models Image Process. vol. 53. pp. 231-239, 1991.

[8] R. Shultz. R. Stevenson. "Extraction of high-resolution frames from video sequences," IEEE Trans. Image Process, pp. 996-1011, 1996.

[9] ITU-R. "Methodology for the subjective assessment of the quality of television pictures." Tech. Rep. BT.500-11. ITU-R, 2002.

[10] A. V. Murthy and L. J. Karam. "IVQUEST- Image and Video Quality Evaluation Software," http://ivulab.asu.edu/Quality/IVQUEST

[11] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol. 13, no. 4, pp. 600-612, Apr. 2004.

[12] H. R. Sheikh and A. C. Bovik, "Image Information and Visual Quality," IEEE Trans. Image Process., pp. 430- 444, Feb. 2006.

[13] D.M. Chandler and S.S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," IEEE Trans. Image Process., vol. 16, no. 9, pp. 2284-2298, Sept. 2007.

[14] Z. Wang and A. C. Bovik, "A Universal Image Quality Index," IEEE Signal Processing Letters, vol. 9, no. 3, pp. 81-84, March 2002.

[15] L. A. Mittal, R. Soundararajan and A. C. Bovik, "Making a Completely Blind Image Quality Analyzer ," IEEE Signal processing Letters, pp. 209-212, vol. 22, no. 3, March 2013.

[16] Xiang Zhu, and Peyman Milanfar, "Automatic Parameter Selection for Denoising Algorithms Using a No-Reference Measure of Image Content," IEEE Trans. Image Process., pp. 3116-3132, Dec. 2010.

[17] N. D. Narvekar and L. J. Karam, "A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD)," IEEE Trans. Image Process., pp. 2678-2683, Sept. 2011.