CNN-BASED PRE-PROCESSING AND MULTI-FRAME-BASED VIEW TRANSFORMATION FOR FISHEYE CAMERA-BASED AVM SYSTEM

*Dong Yoon Choi[1)], Ji Hoon Choi[1)], Jin Wook Choi[2)], Byung Cheol Song[1)]**

[1)]Department of Electronic Engineering, Inha University, Republic of Korea
[2)]Hyundai Motor Company, ADAS Development Team, Republic of Korea
*E-mail: bcsong@inha.ac.kr

ABSTRACT

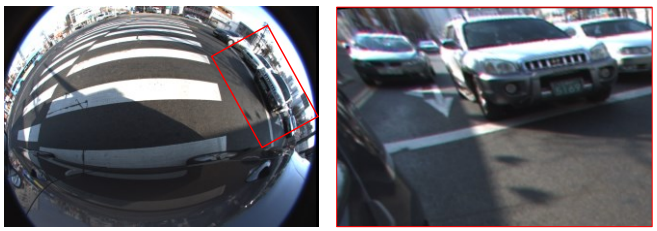
The edges of the wide angle (WA) image generally have poor definition and resolution, which often causes deterioration of the around view monitor (AVM) image quality. This paper proposes a convolutional neural network (CNN)-based preprocessing and a multi-frame-based view transformation to solve this problem, and presents an AVM system based on these methods. First, we analyze the general distortion characteristics of the WA image, and propose a preprocessing using the CNN learning model based on the analysis result. Next, in the view transformation (VT) of the outer edge of the WA image, the inherent problem of low pixel density is solved through motion compensation and hole filling using adjacent frames. Experimental results show that the AVM images by the proposed methods are superior to general AVM images in terms of objective image quality as well as subjective image quality.

Index Terms— Fisheye image, AVM, view transformation, multi-frame, pre-processing, convolutional neural network

1. INTRODUCTION

The fisheye camera or wide angle (WA) camera has a very wide field of view (FOV), and is used in various fields such as surveillance, robot vision, and advanced driver assistance system (ADAS) for automobiles. For example, an automotive around view monitor (AVM) system converts WA images obtained from the fisheye cameras installed at a vehicle into a narrow-angle (NA) image(s) at the top of the vehicle. Fig. 1(a) is an example of a WA image obtained from a WA camera at the right side of a vehicle, and its part can be converted into a NA image in the AVM image generation process (see Fig. 1(b)), e.g., the outer region like the red box. However, since the outer edge of the WA image fundamentally has less sharpness and lower resolution than the center, the visual quality of the converted NA image can be degraded as shown in Fig. 1(b). As a result, this leads to the deterioration of the image quality of the AVM image.

To solve this problem, Choi et al. proposed an AVM system based on a self-example-based edge enhancement method [1], and a learned finite impulse response (FIR) filter-based super-resolution (SR) technique [4]. [1] improved the deteriorated image quality of the WA image more effectively than the unsharp masking method [3] or the deblur [2] method. The FIR-based SR [4] has advantages such as excellent restoration performance in the edge region and possibility of real-time processing owing to its low complexity. However, they have a limitation that an unnatural phenomenon occurs in a complex pattern area such as a texture.



(a) (b)

Fig. 1. Application of fisheye camera in automobile (a) WA image taken from vehicle side (b) NA image corresponding to the red box of Fig. 1 (a).

On the other hand, a few single-image-based SR methods based on convolutional neural network (CNN) [5, 6] have been reported to significantly improve image quality over conventional SR techniques recently. However, CNN-based SR methods have a limited performance improvement when the distortion is severe as in the outer edge region of the WA image. Also, as in [1], a single-image-based AVM system cannot solve the problem of lack of valid pixel information in the view transformation process of the outer frame.

In this paper, we propose CNN-based pre-processing and multi-frame-based view transformation (MFVT) to maximize the sharpness and resolution of view-converted NA images. For the CNN-based preprocessing, we first modeled the degradation of the real-world WA images, and learned the CNN parameters based on the degradation model. MFVT maps pixels from neighboring WA images to a single NA image grid through motion compensation, similar to general multi-frame-based SR (MFSR) techniques [7, 8]. A specific hole filling is finally performed on the remaining holes. Experimental results show that the proposed methods provide superior sharpness and resolution to conventional single-frame-based view transformation

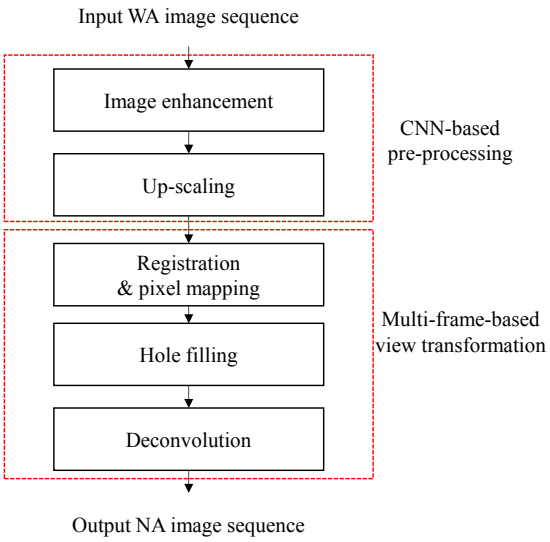


Fig. 2. The overall framework of the proposed algorithm.

(SFVT). Also, the proposed VT shows PSNR result of 2.12dB and SSIM result of 0.0195 higher than the conventional method.

2. THE PROPOSED ALGORITHM

The overall framework of the proposed algorithm is shown in Fig. 2. First, the pre-processing consisting of image enhancement and up-scaling is performed on each frame of the input WA image sequence to improve sharpness and resolution. Next, registration between adjacent frames is performed by motion compensation using optical flow. Then, pixels in the neighboring WA images are mapped to the grid of the reference NA image using a pre-determined look-up-table (LUT). Then, a blurred high-resolution (HR) NA image is generated through multi-frame-based forward warping. Finally, the deblurred HR NA images are obtained through a specific deconvolution process [10].

2.1 CNN-based pre-processing

As shown in Fig. 3(a), the outer region of the WA image is not only blurred but also has jagging artifact. This step consists of image enhancement process and up-scaling process as shown in Fig. 4. The CNN model used here adopts the structure of VDSR [5], which is a CNN-based SR technique for restoring the HR image by estimating the residual component of the input LR image. The parameters of CNN modules for image enhancement and up-scaling are obtained through independent training processes.

Image enhancement
For CNN-based image enhancement, we first synthesize degraded images like Fig. 3(b) from several original resolution label images such as Fig. 3(c). Note that the degradation model results from the optical characteristics of

Fig. 3. (a) A peripheral region of a WA image (b) the degraded image corresponding to an original image (c).

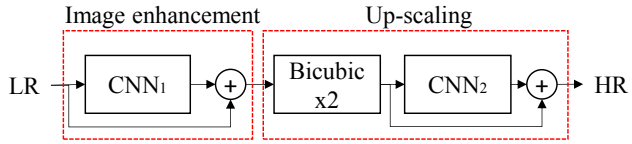


Fig. 4. Block diagram of CNN-based pre-processing process.

the outer regions of the WA images. As shown in Eq. (1), the ground truth image I_{label} is scaled by 1/2 scale (D), then nearest neighbor interpolation is used to enlarge (U). And Gaussian blur is performed to synthesize the input image I_{input} which includes jagging pattern and blur phenomenon.

$$I_{input} = G * U_{NN}(D(I_{label})) \quad (1)$$

Parameters of CNN_1 are trained using a sufficient number of pairs of input data I_{input} and label data I_{label}.

Up-scaling
The up-scaling process adopted the state-of-the-art SR method, i.e., VDSR [5]. For the learning process, low-resolution (LR) images I_{LR} are generated as in Eq. (2). Ground truth I_{HR} is down-scaled by bi-cubic filter, and the down-scaled image is enlarged twice to generate I_{LR}.

$$I_{LR} = U_{bicubic}(D_{bicubic}(I_{HR})) \quad (2)$$

The parameters of the CNN network (CNN_2) for up-scaling are learned by using the obtained I_{HR} - I_{LR} pairs.

2.2 Multi-frame view transformation

The proposed MFVT consists of three steps, as shown in Fig. 2. The pre-processed WA image sequences are input to this stage.

Registration & pixel mapping
First, the inter-frame registration is performed by applying optical flow [11] between adjacent WA images. Next, based on the predefined mapping information between the WA image and the corresponding NA image, pixels in a plurality of WA images are forward-warped to the target NA image grid. Here, the mapping information can be implemented by so-called look-up-table (LUT). The LUT stores position

information in a NA image which each pixel in a WA image is mapped to. Pixels in non-integer positions whose mapping information does not exist in the LUT are interpolated using LUT information corresponding to nearest integer positions.

Hole filling

When pixels in adjacent WA images are mapped to a grid in a NA image, most pixels are mapped to non-integer unit positions due to the precision limit of the LUT. If there is no pixel mapped at a specific integer position, a hole can be generated. So, the hole filling process is required.

The hole filling process can basically be performed by using a distance weight sum with adjacent pixels. The distance weight between the position **x** of the pixel to be estimated and the position **y** of the adjacent pixel mapped to the grid is calculated by Eq. (3).

$$w_{dist}(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\delta_d^2}\right) \tag{3}$$

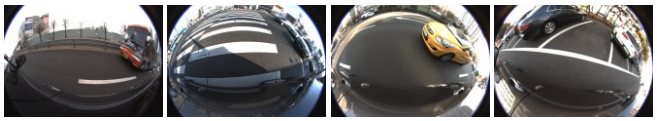
where δ_d indicates the standard deviation of the Gaussian distribution, and $\|\bullet\|$ stands for L_2-norm.

On the other hand, due to the high speed of vehicles and inherent distortion of the WA images some outliers can occur during the registration process. Therefore, if only distance weight is used, artifacts may occur in hole filling result. To compensate for this, motion error-based weight should be defined additionally. The error weight is adjusted according to the magnitude of the motion compensated error in the optical flow process as shown in Eq. (4), so that the weight of the corresponding pixel is reduced in the filtering process.

$$w_{error}(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|I_{ref}(\mathbf{y}) - I_{cur}(\mathbf{x})\|^2}{2\delta_e^2}\right) \tag{4}$$

where $I_{ref}(\mathbf{y})$ means a pixel mapped from a neighboring reference WA image based on a NA image grid, and $I_{cur}(\mathbf{x})$ stands for the corresponding pixel in the current frame. And, δ_e indicates the standard deviation of the Gaussian distribution. Finally, based on the distance weight w_{dist} and the error weight w_{error} the initial HR NA image hole-filled by Eq. (5), i.e., $I_{iniNA}(\mathbf{x})$ is obtained as the bilateral filter [11].

$$I_{iniNA}(\mathbf{x}) = \frac{\sum_{\mathbf{y}\in\Omega} w_{dist}(\mathbf{x},\mathbf{y}) w_{error}(\mathbf{x},\mathbf{y}) I_{ref}(\mathbf{y})}{\sum_{\mathbf{y}\in\Omega} w_{dist}(\mathbf{x},\mathbf{y}) w_{error}(\mathbf{x},\mathbf{y})} \tag{5}$$



(a) (b) (c) (d)

Fig. 5. WA test images. (a) Test1 (b) test2 (c) test3 (d) test4.

where Ω is a set of all possible pixels that are mapped to **x** from the adjacent frames.

Deconvolution

The initial HR NA image may face with blur phenomenon during hole filling. Therefore, it is necessary to apply deconvolution to the initial result like general MFSR approaches. In this paper, we apply the deconvolution process according to Eq. (6) proposed in [10] to $I_{iniNA}(\mathbf{x})$ as the last step, and obtain a final output image I_{outNA}.

$$\min_{I_{outNA}} \lambda \|I_{outNA} \otimes \boldsymbol{k} - I_{iniNA}\|_2^2 + \|\nabla_x I_{iniNA}\|_\alpha + \|\nabla_y I_{iniNA}\|_\alpha \tag{6}$$

In Eq. (6), ∇_x and ∇_y are derivative filters and **k** is the blur kernel. λ and α which indicate the Lagrange multiflier and p-norm were set to the default value used in [10], and the blur kernel was experimented and set to a 5x5 Gaussian filter.

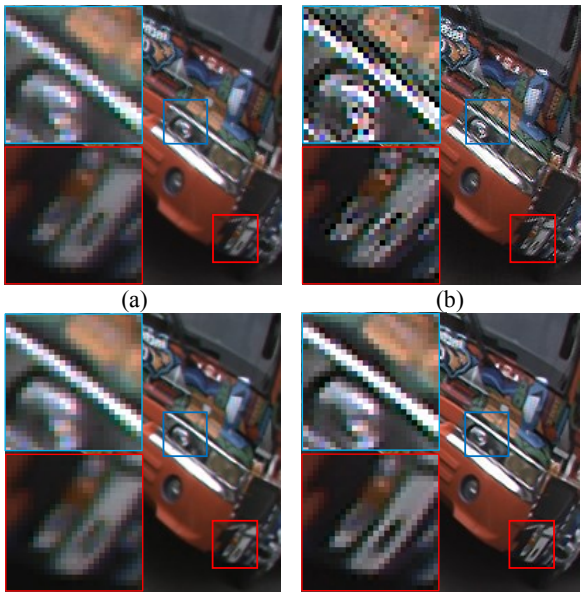
3. EXPERIMENTAL RESULTS

Test video sequences taken in actual driving environment were employed for performance evaluation of the proposed algorithm (see Fig. 6). The GS3-U3-23S56C-C camera from Pointgrey and the fisheye lens called the FE185C057HA-1 from Fujinon were adopted for acquiring the test image sequences. The field of view (FOV) of the image is 185 degrees, the resolution is 1024x720, and the frame rate is 180fps. Our goal is to produce the NA image of the side mirror view from the test WA images. The proposed method is implemented in MATLAB in the following environment: i7-6700k, Geforce Titan X, and RAM of 32Gb.

The evaluation of the proposed method is divided into two parts: CNN-based pre-processing and MFVT. For quantitative evaluation, JNBM (Just Noticeable Blur Metric) [14] for the preprocessing technique was employed and PSNR and SSIM metrics for the MFVT were adopted.

3.1 Performance evaluation of CNN-based pre-processing

Since the SR part in CNN-based pre-processing adopted VDSR as it is, only the performance of CNN-based image enhancement was evaluated in this section. The CNN structure consists of 20 layers with 64 depths, as in VDSR[5], and a 3x3 convolution filter is used. The CNN parameters were learned using MatConvNet [13], and the 91 image sets [12] were adopted as training images.



(a) (b) (c) (d)

Fig. 6. Pre-processed results for Fig. 5(a) (a) Input WA image (b) deblur [2] (c) self-example-based edge enhancement [1] and (d) the proposed CNN-based method.

For the purpose of improving the WA image, we compared the proposed algorithm with the self-example-based method [1] and Xu's deblur method [2]. Fig. 6 compares the pre-processed results for Fig. 5 (a). For convenience of comparison, the blur area (red box) of the input WA image (Fig. 6 (a)) and the blue box area where the jagging artifact exists are mainly compared. Xu's [2] deblur technique improves clarity but tends to boost artifacts. The self-example-based edge enhancement [1] overcomes jagging artifacts, but seldom works for complex patterns such as red box. However, the proposed preprocessing successfully improves the sharpness without artifacts.

On the other hand, Table I shows that the proposed CNN technique shows 0.9754 and 0.8950 higher JNBM than the input image and [1], respectively. Xu's [2] shows higher JNBM than proposed method, but this is because artifacts are severely emphasized as shown in Fig. 6.

3.2 Performance evaluation of multi-frame-based view transformation

The proposed MFVT was compared with a conventional backward warping based SFVT. The two parameters for the hole filling, i.e., δ_d and δ_e were set to 0.75 and 40, respectively. Also, the number of adjacent frames was set to 9 frames. Seeing Fig. 7(a) SFVT provides a limited quality even when pre-processing was used together. However, MFVT shows sharp edges and greatly improved resolution. Note that the plate number is clearly visible in Fig. 7(b).

On the other hand, the restoration performance of synthetic images is compared for the quantitative evaluation of the proposed method.

(a) (b)

Fig. 7. VT results for Fig. 5(b) (a) SFVT (b) MFVT.

Table I JNBM results of several algorithms

	Input	Deblur[2]	Self-ex[1]	Proposed
Test1	6.9441	8.5026	6.8438	7.7942
Test2	6.4282	7.8181	6.5297	7.2716
Test3	6.1603	7.6616	6.1826	6.9280
Test4	6.3005	8.5008	6.5987	7.7410
Average	**6.4583**	**8.1208**	**6.5387**	**7.4337**

Table II Comparison of SFVT and MFVT

	PSNR [dB]		SSIM	
	SFVT	MFVT	SFVT	MFVT
Bigships	34.35	36.37	0.9224	0.9406
City	32.37	34.87	0.9219	0.9511
Shields	31.22	33.34	0.9112	0.9259
Mobcal	31.44	33.29	0.9129	0.9288
Average	**32.35**	**34.47**	**0.9171**	**0.9366**

WA images were synthesized by inverse transformation of four well-known 720p images such as *bigships, city, shields,* and *mobcal*. Note that in this experiment, only the pure VT process was compared without pre-processing and deconvolution steps. Table II shows that MFVT provides higher PSNR of 2.1dB and higher SSIM of 0.0195 than SFVT.

4. CONCLUDING REMARKS

In this paper, we propose CNN-based pre-processing and MFVT for generating HR NA images from input WA images. First, CNN-based pre-processing improves the clarity and resolution of the input WA images prior to view transformation. Second, MFVT improves the definition of the NA image by compensating for the insufficiency of the pixel density in the outer edge of the WA images. As a future work, we will develop a common CNN framework that integrates image enhancement and upscaling, and replace current motion compensation based hole filling with CNN-based approach.

5. ACKNOWLEDGMENT

This work was supported by the Industrial Strategic Technology Development Program (10073154, Development of human-friendly human-robot interaction technologies using human internal emotional states recognition) funded by the Ministry of Trade, industry & Energy(MI, Korea) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2016R1A2B4007353)

6. REFERENCES

[1] D. Y. Choi, J. H. Choi, J. W. Choi, and B. C. Song, "Self-example-based edge enhancement algorithm for around view monitor images," *Proc. IS&T International Symposium on Electronic imaging,* 2017.

[2] L. Xu , S. Zheng and J. Jia, "Unnatural l0 sparse representation for natural image deblurring," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.1107-11142, 2013.

[3] A. Polsel, G. Ramponi and V. J. Mathews, "Image enhancement via adaptive unsharp masking," *IEEE Transaction on Image Processing*, vol. 9, no. 3, pp. 505-510, 2000.

[4] D. Y. Choi and B. C. Song, "Fast super-resolution algorithm using ELBP classifier," *Proc. IEEE Visual Communications and Image Processing*, Dec. 2015.

[5] J. Kim, J. K. Lee, K and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," *Proc. IEEE Computer Vison and Pattern Recognition*, 2016.

[6] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295-307, 2016.

[7] E. Faramarzi, D. Rajan, F. Fernadaes, M. Christensen, "Blind super-resolution of real-life video sequences," *IEEE Transactions on Image Processing*, vol. 25 no. 4, pp. 1544-1555, 2016.

[8] M. Bätz, A. Eichenseer, and A. Kaup, "Multi-image super-resolution using a dual weighting scheme based on Voronoi tessellation," *Proc. IEEE International Conference on Image Processing,* 2016.

[9] C. Tomasi, and R. Manduchi, "Bilateral filtering for gray and color images," *Proc. IEEE International Conference on Computer Vision*, 1998.

[10] D. Krishnan, T. Tay, and R. Fergus, "Blind deconvolution using a normalized sparsity measure," *Proc. IEEE Computer Vision and Pattern Recognition,* 2011.

[11] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," *Doctoral Thesis. Massachusetts Institute of Technology,* 2009.

[12] J. Yang, J. Wright, T. S. Huang and M. Yi, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861-2873, Nov. 2010.

[13] A. Vedaldi, and K. Lenc, "Matconvnet: convolutional neural networks for MTALB," *Proc. ACM International Conference on Multimedia,* 2015.

[14] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Transaction on Image Processing*, vol. 18, no. 4, pp. 717-728, 2009.