Exemplar-based Human Body Super-resolution for Surveillance Camera Systems

Kento Nishibori[1], Tomokazu Takahashi[1,2], Daisuke Deguchi[3], Ichiro Ide[1] and Hiroshi Murase[1]
[1]*Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan*
[2]*Faculty of Economics and Information, Gifu Shotoku Gakuen University, Nakauzura 1-38, Gifu-shi, Gifu-ken, 500-8288, Japan*
[3]*Information and Communications Headquarters, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan*

Keywords: Exemplar-based Super-resolution, Human Body Image, High-frequency Component, Surveillance System, Image Quality Assessment.

Abstract: In this paper, we propose an exemplar-based super-resolution method applied to a human body in a surveillance video. Since persons are usually captured as low-resolution images by a video surveillance system, it is sometimes necessary to perform detection and identification of persons from not only a human face but also from the human body appearance. The super-resolution for a human body image is difficult because the appearances of person images vary according to the color of clothing and the posture of persons. Thus, we focus on the high-frequency components that could restore the lost high-frequency components of the low-resolution image regardless to the variation of the clothing. Therefore, the purpose of the work presented in this paper is to apply the exemplar-based super-resolution using high-frequency components for a low-resolution human body image to generate a high-resolution human body image so that both computer systems and humans can identify persons more accurately. As a result of experiments, we confirmed the effectiveness of the proposed super-resolution method.

1 INTRODUCTION

In recent years, many video surveillance systems are installed increasingly for preventing crime and terrorism at various public places such as airports, railway stations, streets, and buildings. However, it becomes more difficult for human operators to detect terror suspects in proportion to the increase. Additionally, as the video surveillance usually monitors a wide area, persons are usually captured as low-resolution (LR) images.

The super-resolution (SR) is a promising method for enhancing the quality of LR image by compensating high-frequency components (Bonet, 1997; Lin and Shum, 2004; Wang et al., 2009; Zeyde et al., 2010; Milanfar, 2011; Jiang et al., 2012b; Ho and Zeng, 2012), and several researches on its application to facial images have been performed. Since the positions and the shapes of face parts are able to be estimated stochastically, they are used as clues for face image SR (Baker and Kanade, 2000; Baker and Kanade, 2002; Freeman et al., 2002; Liu et al., 2007; Jiang et al., 2012a; Yoshida et al., 2012; Ma et al., 2013). However, since it is difficult to capture a human face appropriately using a surveillance video, it is sometimes necessary to perform detection and identification of persons from the human body appearance (Nakajima et al., 2003). On the other hand, the SR for human body images is difficult since the appearance of a human body has a large variation such as body shapes, postures, and clothing. Therefore, we can not simply apply the face SR method to human body images.

The aim of the work presented in this paper is to generate a human body image in LR to a high-resolution (HR) image for enabling both a human and a computer system to conduct the identification process more accurately. Among various problems, in this paper, we propose a method for exemplar-based SR of human body images specifically focusing on the problem of variation of clothing.

Nishibori K., Takahashi T., Deguchi D., Ide I. and Murase H..
Exemplar-based Human Body Super-resolution for Surveillance Camera Systems.
DOI: 10.5220/0004686101150121
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 115-121
ISBN: 978-989-758-003-1

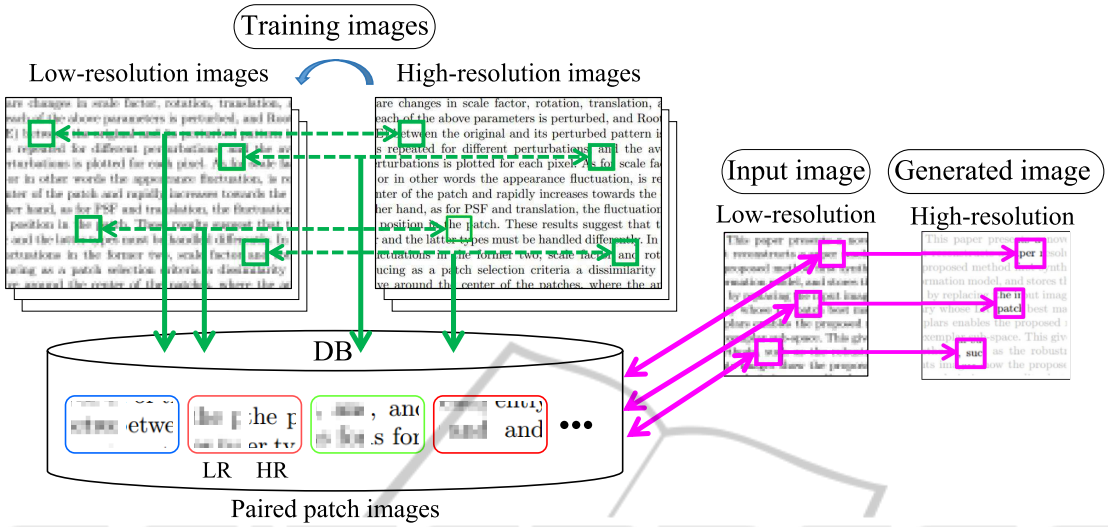


Figure 1: The framework of an existing exemplar-based super-resolution method (Shibata et al., 2013).

2 SUPER-RESOLUTION FOR A BODY IMAGE

2.1 Existing Method: Exemplar-based Super-resolution

Figure 1 shows the generation process of an existing SR image with the exemplar-based method (Shibata et al., 2013), which we use in our method as a basis. The procedure of the SR method is as follows:

(i) HR training images are downsampled by a factor of $1/r$ to obtain LR training images. HR and LR paired patches are extracted at the size of $rL \times rL$ [pixels] and $L \times L$ [pixels] from HR and LR images respectively, preserving their positional relationship. These paired patches are stored in a database.

(ii) Then, patches extracted from the LR input image at the size of $L \times L$ [pixels] are matched up with LR patches in the database, and the most similar LR patches in the database are selected.

(iii) The SR image is generated by replacing the LR input patch images with the HR patches corresponding to the selected LR patches.

However, this exemplar-based super-resolution is not suitable for human body images because it is difficult to create a database which covers the variation of the human body appearance.

2.2 Proposed Method: Exemplar-based Super-resolution using High-frequency Components

The exemplar-based super-resolution method can provide superior performance when LR patches extracted from an input image sufficiently match the LR and HR paired patches in the database. Thus, it is necessary to prepare a rich dataset which covers the variation of appearances in respect to LR input images. However, since the appearances of person images vary according to the color of clothing and the posture of persons as shown in Figure 2(a), we need a large number of examples and also it requires a time-consuming process.

Thus, we focused on the high-frequency components that could restore the lost high-frequency components in a low-resolution image regardless to the variation of the appearance. Figure 2(b) displays the high-frequency components of the body images in Figure 2(a), where we can see that the texture of the clothing are very similar in spite of their different colors. The high-frequency components are obtained by the difference between HR and LR images, where the HR image is downsampled to generate LR images by a factor of $1/r$. In order to generate an HR image, high-frequency components of the training images are applied to restore the lost high-frequency components of the LR images.

Figure 3 shows the proposed SR method using the high-frequency components. The procedure is as follows:

(i) An HR image $I_k^{(0)}$ is downsampled to create an LR

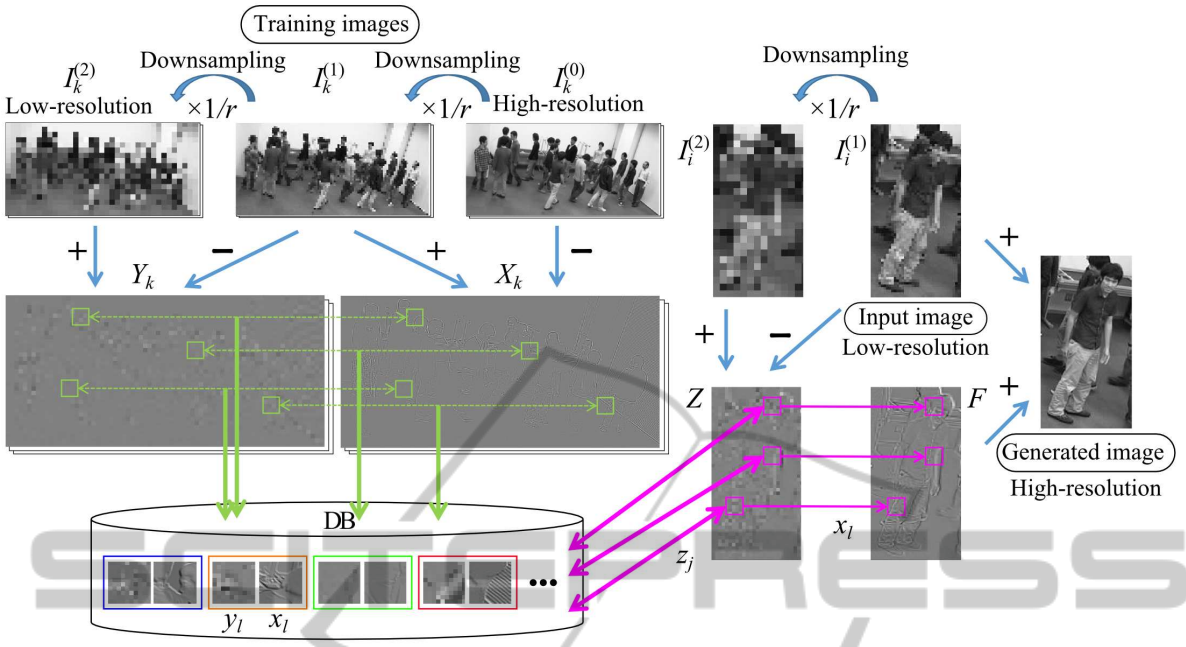
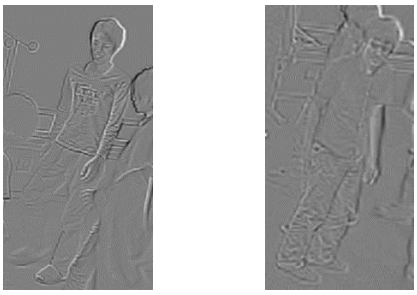


Figure 3: Framework of the proposed super-resolution method.

(a) Body images with clothing.



(b) High-frequency components of body images.

Figure 2: High-frequency components in different clothing.

image $I_k^{(1)}$ by a factor of $1/r$. The LR image is also downsampled to generate an even low-resolution (LR') image $I_k^{(2)}$ by a factor of $1/r$ which corresponds to downsampling HR image $I_k^{(0)}$ by a factor of $1/r^2$. The high-frequency components X_k are obtained by the difference between an HR image $I_k^{(0)}$ and an LR image $I_k^{(1)}$, as follows.

$$X_k = I_k^{(1)} - I_k^{(0)}$$
$$I_k^{(1)} = D(I_k^{(0)}) \quad (1)$$

The high-frequency components Y_k are also obtained by the difference between an LR image $I_k^{(1)}$ and an LR' image $I_k^{(2)}$, as follows.

$$Y_k = I_k^{(2)} - I_k^{(1)}$$
$$I_k^{(2)} = D(I_k^{(1)}) \quad (2)$$

Here, I_k represents the k-th training image, and $D(\cdot)$ represents the downsampling process.

(ii) Paired patches x_l and y_l are extracted at the size of $rL \times rL$ [pixels] and $L \times L$ [pixels] from X_k and Y_k preserving the positional relationship between the X_k and Y_k components. Here, x_l and y_l represent the l-th patches extracted from X_k and Y_k respectively. These paired patches x_l and y_l are extracted at the size of $rL \times rL$ [pixels] and $L \times L$ [pixels] respectively, and stored in a database to create the training dataset.

(iii) An LR input image $I_i^{(1)}$ is downsampled by a factor of $1/r$ to generate an even low-resolution (LR') input image $I_i^{(2)}$, and high-frequency components Z are obtained by the difference between

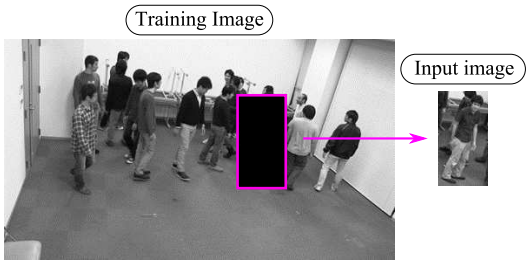


Figure 4: Training image and input image.

LR input image $I_i^{(1)}$ and LR' input image $I_i^{(2)}$, as follows.

$$Z = I_i^{(2)} - I_i^{(1)}$$
$$I_i^{(2)} = D(I_i^{(1)}) \quad (3)$$

In order to simplify the experimental condition, the LR input image $I_i^{(1)}$ is generated by downsampling an original HR input image $I_i^{(0)}$ by a factor of $1/r$, in advance. The patch z_j extracted from the high-frequency components Z are matched up with patches y_l extracted from Y_k in the database, and the most similar patches in the database are selected.

(iv) The high-frequency components F is generated by replacing the patch z_j extracted from the high-frequency components Z with the patch x_l corresponding to the selected patch y_l. Finally, the high-frequency components F and the LR input image are combined to generate the SR image.

3 EXPERIMENT

3.1 Experimental Conditions

We captured ten HR images with a Canon video camcorder iVIS HF G10 at a resolution of 1,920 × 1,080 pixels. Each image contained a group of 17 people under the same illumination condition. For input data, we manually extracted region images of a person from an HR image and downsampled the extracted image by a factor of $1/r$ = 1/3. For the HR-LR training image database, we extracted patches from the region excluding the input person region as shown in Figure 4. The database consisted of two million pairs of HR and LR patch images.

As the search function for matches in the database, we applied the random kd-tree for approximate nearest neighbor search. For this, we used FLANN (Fast Library for Approximate Nearest Neighbors) from the OpenCV library (Muja and Lowe, 2009).

We applied two different conditions on patches for the evaluation of SR images; one with luminance and one with high-frequency components. As an image quality assessment, we used the structural similarity (SSIM) (Wang et al., 2004), which can evaluate the similarity between a reference image and a distorted image from the point of human visual perception better than other image quality metrics such as the mean squared error (MSE) or the peak-signal-to-noise ratio (PSNR).

3.2 Evaluation of the Image Quality of the Generated Images

In order to confirm the validity of the proposed SR method, we compared the quality of images between the original image, an LR input image, and HR images obtained by different magnification methods which are the bi-cubic interpolation, and the existing SR method using luminance, and the proposed SR method using high-frequency components.

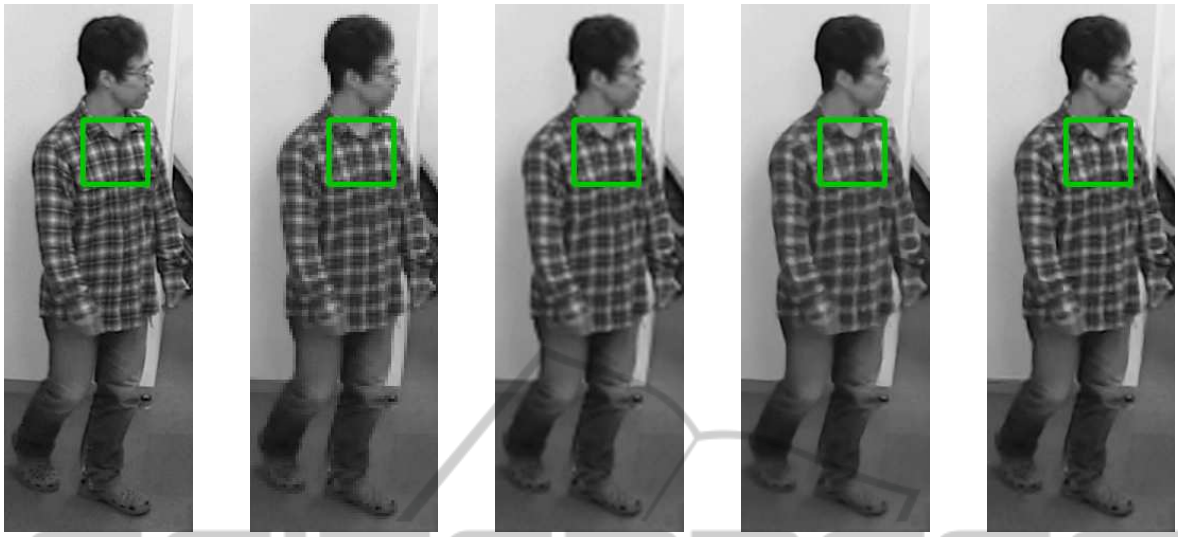
Figure 5 shows examples of an LR input image and HR images obtained by different magnification methods. In this experiment, LR images were magnified by a factor of r = 3, and SR was performed when patch images were extracted from LR images at the size of 7 × 7 pixels. Figure 5(a) shows human full-body images, and Figure 5(b) shows partial zoom-ups of the full-body images for comparing the detail of the texture. Figure 5(c) shows the SSIM maps. Figure 5(a)-(i) indicates the original HR images, which are used as reference in the SSIM image quality evaluation. The size of images in Figure 5(a) was 178 × 499 pixels, Figure 5(a)-(i) was downsampled by a factor of $1/r$ = 1/3 to generate an LR input image (59 × 166 pixels) for the SR.

Figure 5(a)-(ii) was magnified using the nearest neighbor interpolation by a factor of r = 3, and 5(a)-(iii) was magnified using the bi-cubic interpolation. Figure 5(a)-(iv) was magnified using the luminance components. Figure 5(a)-(v) was magnified using the high-frequency components. We could have normalized each patch to make it robust against change in lighting conditions, but we did not do so because there was not much change between the training images and the input images in this experiment.

Table 1 shows the comparison of SSIM by different magnification methods. We can see that the qual-

Table 1: Comparison of SSIM by different magnification methods.

Method	LR input	Bi-cubic	Luminance	HF
SSIM	0.855	0.900	0.912	0.928



(i) Original (ii) Low-resolution input SSIM = 0.856 (iii) Bi-cubic SSIM = 0.907 (iv) Luminance (Existing SR) SSIM = 0.887 (v) High-frequency (Proposed SR) SSIM = 0.912

(a) Human body images.

(b) Zoom-ups of part of the full-body images.

(c) SSIM maps in image quality evaluation.

Figure 5: Example of high-resolution images obtained by different methods.

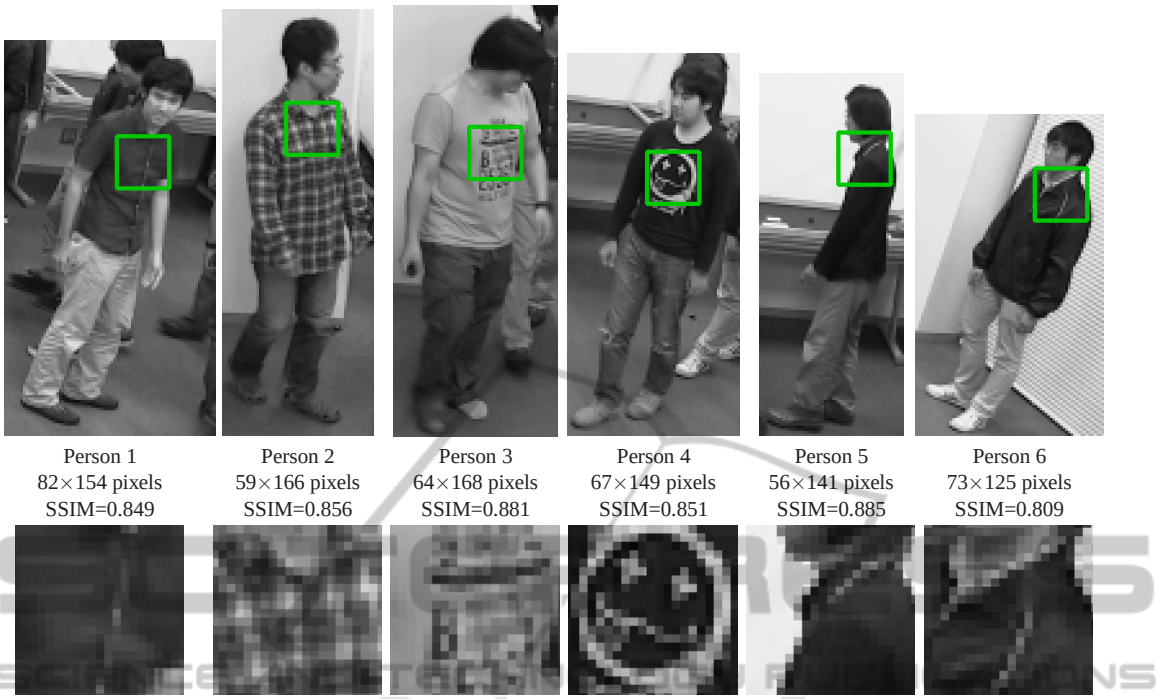
ity of the SR image using high-frequency components surpasses the other images.

Figure 6(a) shows the body image of six persons, and Figure 6(b) shows the SR images using the proposed method. Figure 7 shows the average of SSIM when the body image of six persons were magnified by a factor of $r = 3$, and patch images were extracted from both training and input images at the size of 3×3 pixels to 21×21 pixels. When luminance was used as the feature, we can see that the image quality deteriorates according to the patch size. Meanwhile, we can see that when using high-frequency components, it turns out that there is little influence on the image quality according to the patch size.

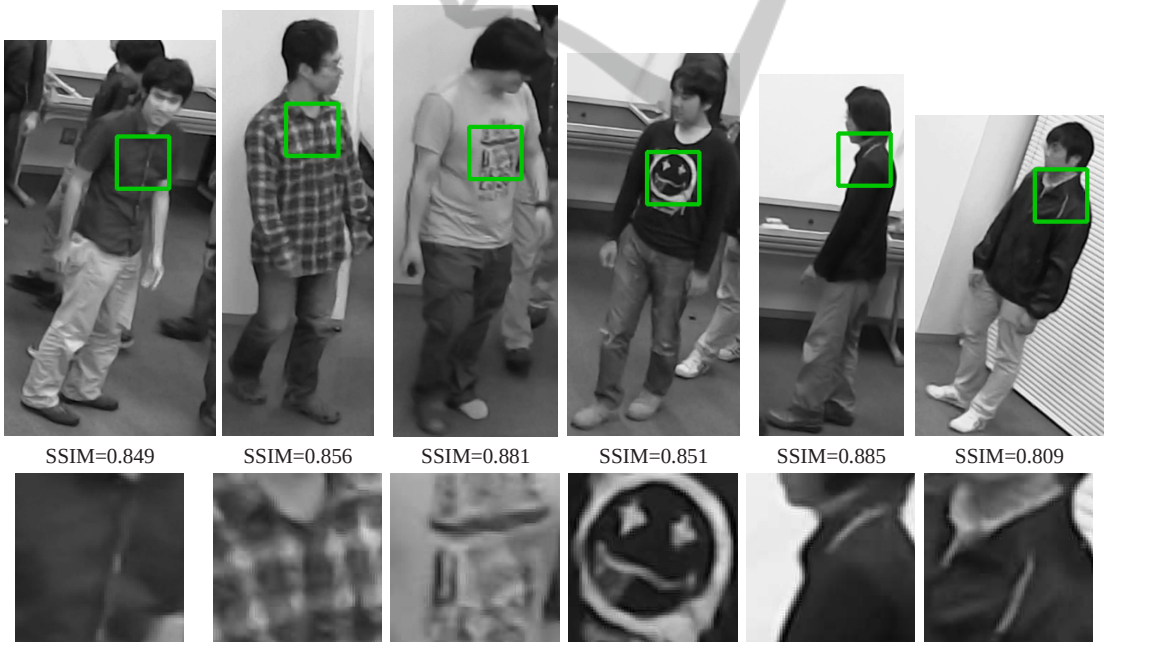
Figure 8 shows the average of SSIM when the body images of six persons were magnified by factors of ($r =$) 2 to 5. Even if the magnification factor is changed, we can see that the quality of the SR images using high-frequency components are better than that using bi-cubic interpolation.

4 CONCLUSIONS

In order to perform SR of human body images, we proposed the exemplar-based SR using the high-frequency components of human body images. As a result of experiments, the quality of the magnified image using the high-frequency components of training images surpassed the comparative methods. We



(a) Low-resolution input images.



(b) Super-resolution images by the proposed method.

Figure 6: Examples of high-resolution images obtained by the proposed method.

also confirmed that in different magnification factors of (r =) 2 to 5, the proposed method also surpassed the comparative methods.

As future work, there are two challenges. First, we will improve the proposed exemplar-based SR method so that it generate images more accurately and robustly by adopting time series information. Secondly, we will actually apply the method to surveillance video captured in real environments, such as airports, stations, streets, and buildings.

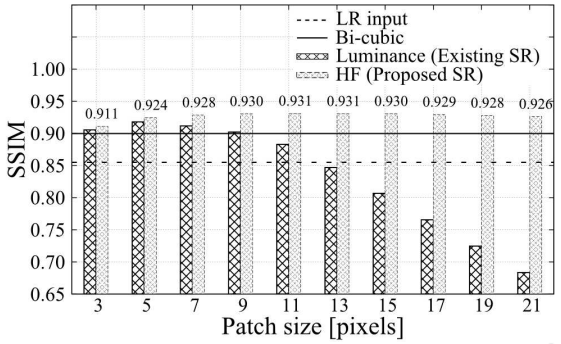


Figure 7: SSIM for different patch sizes.

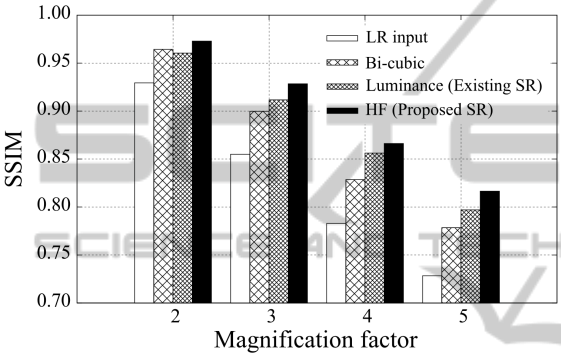


Figure 8: SSIM for different magnification factors.

ACKNOWLEDGEMENTS

This work was supported by the "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society," Special Coordination Fund for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government. The authors wish to thank the members of Murase laboratory participating in the video recording used in the experiment.

REFERENCES

Baker, S. and Kanade, T. (2000). Hallucinating faces. In *Proc. IEEE Fourth Int'l Conf. Automatic Face and Gesture Recognition (FG'00)*, pages 83–88.

Baker, S. and Kanade, T. (2002). Limits on super-resolution and how to break them. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(9):1167–1183.

Bonet, J. S. D. (1997). Multiresolution sampling procedure for analysis and synthesis of texture images. In *Proc. ACM 24th Int'l Conf. Computer Graphics and Interactive Techniques (SIGGRAPH'97)*, pages 361–368.

Freeman, W. T., Jones, T. R., and Pasztor, E. C. (2002). Example-based super-resolution. *IEEE Trans. Computer Graphics and Applications*, 22(2):56–65.

Ho, T. and Zeng, B. (2012). Super-resolution image by curve fitting in the threshold decomposition domain. *Visual Communication and Image Representation*, 23(1):208–221.

Jiang, J., , Hu, R., Han, Z., Lu, T., and Huang, K. (2012a). Position-patch based face hallucination via locality-constrained representation. In *Proc. IEEE 13th Int'l Conf. on Multimedia and Expo (ICME'12)*, pages 212–217.

Jiang, J., Hu, R., Han, Z., Huang, K., and Lu, T. (2012b). Efficient single image super-resolution via graph embedding. In *Proc. IEEE 13th Int'l Conf. on Multimedia and Expo (ICME'12)*, pages 610–615.

Lin, Z. and Shum, H. Y. (2004). Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(1):83–97.

Liu, C., Shum, H. Y., and Freeman, W. T. (2007). Face hallucination: Theory and practice. *ACM Trans. Computer Vision*, 75(1):115–134.

Ma, X., Li, W., Xu, H., Yang, X., and Song, H. (2013). A general residue compensation framework of learning-based face super-resolution. *Computational Information Systems*, 9(10):4049–4056.

Milanfar, P. (2011). *Super-Resolution Imaging (Digital Imaging and Computer Vision)*. CRC Press.

Muja, M. and Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. In *Proc. Fourth Int'l Conf. Computer Vision Theory and Applications (VISSAP'09)*, pages 331–340.

Nakajima, C., Pontil, M., Heisele, B., and Poggio, T. (2003). Full-body person recognition system. *Trans. Pattern Recognition in Kernel and Subspace Methods for Computer Vision*, 36(9):1997–2006.

Shibata, T., Iketani, A., and Senda, S. (2013). Single image super resolution reconstruction in perturbed exemplar sub-space. In *Proc. IEEE 12th Conf. Asian Conference on Computer Vision (ACCV'13)*, pages 401–412.

Wang, J. T., Liang, K. W., Chang, S. F., and Chang, P. C. (2009). Super-resolution image with estimated high frequency compensated algorithm. In *Proc. 9th Int'l Symp. Communications and Information Technology (ISCIT'09)*, pages 175–180.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612.

Yoshida, T., Takahashi, T., Deguchi, D., Ide, I., and Murase, H. (2012). Robust face super-resolution using free-form deformations for low-quality surveillance video. In *Proc. IEEE 13th Int'l Conf. on Multimedia and Expo (ICME'12)*, pages 368–373.

Zeyde, R., Elad, M., and Protter, M. (2010). On single image scale-up using sparse representation. In *Proc. 7th Int'l Conf. Curves and Surfaces*, pages 711–730.