2012 IEEE International Conference on Multimedia and Expo

Robust Face Super-Resolution Using Free-Form Deformations For Low-Quality Surveillance Video

Tomonari Yoshida*, Tomokazu Takahashi†, Daisuke Deguchi‡, Ichiro Ide* and Hiroshi Murase*
*Graduate School of Information Science, Nagoya University, Japan
yoshidat@murase.m.is.nagoya-u.ac.jp, {ide, murase}@is.nagoya-u.ac.jp
†Faculty of Economics and Information, Gifu Shotoku Gakuen University, Japan
ttakahashi@gifu.shotoku.ac.jp
‡Information and Communication Headquarters, Nagoya University, Japan
ddeguchi@nagoya-u.jp

***Abstract*—Recently, the demand for face recognition to identify persons from surveillance video cameras has rapidly increased. Since surveillance cameras are usually placed at positions far from a person's face, the quality of face images captured by the cameras tends to be low. This degrades the recognition accuracy. Therefore, aiming to improve the accuracy of the low-resolution-face recognition, we propose a video-based super-resolution method. The proposed method can generate a high-resolution face image from low-resolution video frames including non-rigid deformations caused by changes of face poses and expressions without using any positional information of facial feature points. Most existing techniques use the facial feature points for image alignment between the video frames. However, it is difficult to obtain the accurate positions of the feature points from low-resolution face images. To achieve the alignment, the proposed method uses a free-form deformation method that flexibly aligns each local region between the images. This enables super-resolution of face images from low-resolution videos. Experimental results demonstrated that the proposed method improved the performance of super-resolution for actual videos in terms of both image quality and face recognition accuracy.**

***Keywords*-video-based super-resolution; low-resolution face image; free-form deformation; face recognition**

I. Introduction

Recently, a large number of surveillance video cameras are installed in banks, subways, and many other places. The demand for techniques to identify persons from these cameras has rapidly increased for the purpose of criminal investigations and anti-terrorism measures.

Face recognition is an effective approach to identify a person from surveillance video cameras. However, the quality of face images captured by the cameras tends to be low as shown in Figure 1, because the installation position and the number of surveillance cameras are restricted, and also their positions are far from faces. This degrades the recognition accuracy. Use of super-resolution techniques [1]–[8] could be one solution for this problem. There are two types of approaches for super-resolution; single-image-based super-resolution and video-based super-resolution. Face Hallucination methods [1]–[4] belong to the single-image-based approach, which generates a high-resolution image from one low-resolution input image by using a dataset consisting of images of faces other than the input one. The approach requires a large amount of face images in order to learn the relationship between low-resolution face images and high-resolution face images. On the other hand, the video-based approach generates a high-resolution image by complementing information of pixels using the displacement between multiple low-resolution video frames caused by slight changes of face poses and expressions. Since this approach does not require any face image dataset, in this paper, we take the video-based super-resolution approach.

Wheeler et al. [5] have proposed a video-based method that performed an alignment between face images by using an Active Appearance Model (AAM). Meanwhile, Mortazavian et al. [6] used a three-dimensional model of a face for the alignment. These methods require positional information of facial feature points to align face images including non-rigid deformations caused by changes of face poses and expressions. However, for low-resolution images, it is difficult to obtain the accurate positions of the feature points.

Therefore, we propose a video-based super-resolution method for low-resolution face images, which performs super-resolution without using any information of facial feature points. To achieve this, we employ a Free-Form Deformation (FFD) method to align the face images including the non-rigid deformations.

The remainder of the paper is organized as follows. We first describe the detail of the proposed method in Section 2.

Figure 1. Examples of low-quality images from surveillance cameras.

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.162

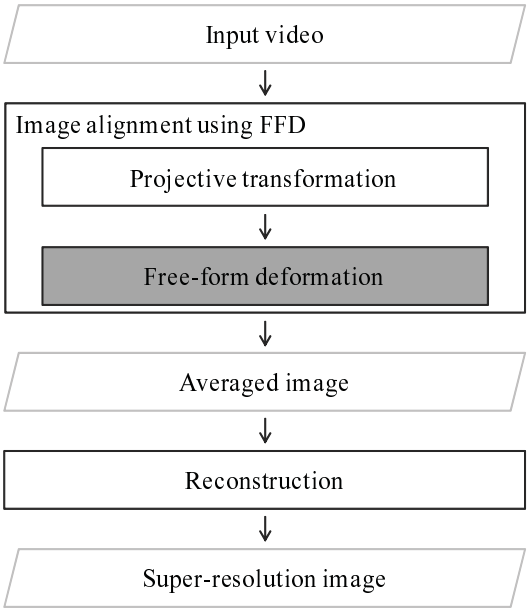

Figure 2. The process flow of the proposed method.

Section 3 presents experimental results, and then we discuss them in Section 4. Finally, we provide conclusions and future work in Section 5.

II. Super-resolution using FFD

Figure 2 shows the flow of the proposed super-resolution method. The proposed method is composed of two processes; an image alignment process and a reconstruction process. The image alignment process performs image alignment with a sub-pixel accuracy between frames in the input video without any facial feature points, and then generates an averaged image by averaging the aligned images. We employ a FFD method [9] to achieve the accurate alignment of low-resolution face images including non-rigid deformations caused by changes of face poses and expressions. The averaged image is a high-resolution image whose pixel values are interpolated from multiple low-resolution images. The reconstruction process estimates a final super-resolution image by removing the optical blur from the averaged image.

A. Image alignment using FFD

The alignment process first chooses one frame from the input video as a reference image for the alignment, and then aligns other frames to the reference image. It is difficult to align images with global displacements, such as two-dimensional rotation or parallel translations by using only the FFD method. Therefore, we take a two-step alignment approach; a projective transformation step for a coarse alignment for the global displacements, and then a FFD step for a precise alignment for local non-rigid deformations caused by changes of face poses and expressions.

Projective transformation

The Inverse Compositional Image Alignment (ICIA) method [10] is used to obtain a projective transformation matrix between images. This method performs the alignment between the reference image and a target image by iteratively updating the transformation matrix to reduce the difference between the reference image and the image transformed from the target image by the matrix.

Free-form deformations

The detailed alignment is performed by using the FFD method. The FFD method has been used for image alignment in various fields, such as medical imaging. This method could flexibly deform an image by moving control points placed on it.

Figure 3 shows the process flow of the FFD method. The alignment between the reference image $\boldsymbol{T}$ and the target image $\boldsymbol{I}$ is performed by the following steps:

Step 1 Put $L \times M$ control points $\boldsymbol{p}_{l,m}$ ($l = 1, \cdots, L$, $m = 1, \cdots, M$) on the target image $\boldsymbol{I}$ in a reticular pattern.

Step 2 Update each $\boldsymbol{p}_{l,m}$ by calculating the following equations:

$$\boldsymbol{p}_{l,m} = \boldsymbol{p}_{l,m} + \mu \frac{\nabla \boldsymbol{c}}{||\nabla \boldsymbol{c}||}, \quad (1)$$

$$\nabla \boldsymbol{c} = \frac{\partial}{\partial \boldsymbol{p}_{l,m}} \sum_{D_{l,m}} |\boldsymbol{I}'(x,y) - \boldsymbol{T}(x,y)|, \quad (2)$$

where $D_{l,m}$ is a set of pixels which exist in the neighborhood of the control point $\boldsymbol{p}_{l,m}$. $\boldsymbol{I}'(x,y)$ and $\boldsymbol{T}(x,y)$ are pixel values of a location (x,y) in $\boldsymbol{I}'$ and $\boldsymbol{T}$, respectively.

Step 3 Obtain a deformed image $\boldsymbol{I}'$ from $\boldsymbol{I}$ by moving the control points using the following equations:

$$\boldsymbol{I}'(x,y) = \boldsymbol{I}(\boldsymbol{w}(x,y)), \quad (3)$$

$$\boldsymbol{w}(x,y) = \sum_{i=0}^{3} \sum_{j=0}^{3} B_i(u') B_j(v') \boldsymbol{p}_{u+i,v+j}, \quad (4)$$

$$B_0(t) = (1-t)^3/6, \quad (5)$$

$$B_1(t) = (3t^3 - 6t^2 + 4)/6, \quad (6)$$

$$B_2(t) = (-3t^3 + 3t^2 + 3t + 1)/6, \quad (7)$$

$$B_3(t) = t^3/6, \quad (8)$$

where $u = \lfloor x/L \rfloor - 1$, $v = \lfloor y/M \rfloor - 1$, $u' = x/L - \lfloor x/L \rfloor$, and $v' = y/M - \lfloor y/M \rfloor$.

Step 4 Calculate the distance d between the deformed image $\boldsymbol{I}'$ and the reference image $\boldsymbol{T}$ by

$$d = \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} |\boldsymbol{I}'(x,y) - \boldsymbol{T}(x,y)|. \quad (9)$$

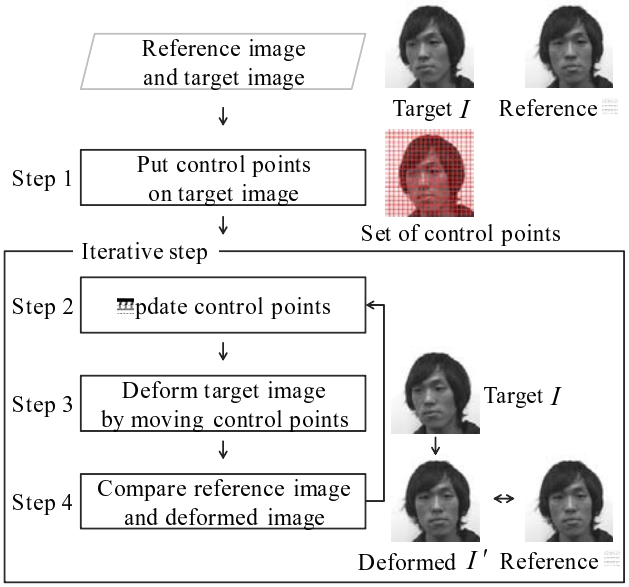

Figure 3. The process flow of the image alignment using FFD.

Step 5 Iterate *Step* 2∼*Step* 4 until d becomes lower than a threshold, or the number of iterations reaches a maximum count. Here, the threshold and the maximum count are defined beforehand.

These steps are applied to each frame in the input video to obtain face images that are aligned to the reference image with a sub-pixel accuracy. After this, the proposed method generates a high-resolution image as an averaged image by averaging the aligned images.

B. Reconstruction

To remove the optical blur from the averaged image, the Maximum A Posteriori (MAP) method [7] is applied in the reconstruction process. The super-resolution image is obtained by minimizing a cost function which is derived from the posterior probability of the image.

The cost function for the MAP method is represented by the following equation:

$$J = \sum_{x=1}^{X'} \sum_{y=1}^{Y'} \Big[\boldsymbol{b}(x,y)^{\mathrm{T}} \boldsymbol{H} - \boldsymbol{R}(x,y) \Big]^2 + \lambda D(\boldsymbol{H}), \quad (10)$$

where $\boldsymbol{H}$ is the super-resolution image ($X' \times Y'$ pixels) corresponding to the averaged image $\boldsymbol{R}$, and $\boldsymbol{b}(x,y)$ is a Point Spread Function (PSF) kernel that expresses the optical blur for each coordinate (x,y). We assume that the PSFs could be represented as the same Gaussian distribution for all coordinates although we can use kernels with various shapes depending on the capturing conditions. λ is a parameter to control the influence of the second term. The first term represents the estimated error of the super-resolution image. Meanwhile, the second term is a constraint term based on the prior probability of the image. We use the regularization term of Bilateral Total Variation (BTV) method [8] as $D(\boldsymbol{H})$, which represents the constraint for the smoothness of the edges. The final super-resolution image is obtained by updating $\boldsymbol{H}$ until J meets the minimum. This minimization is conducted by a conjugate gradient method. The gradient of J is represented as the following equation:

$$\begin{aligned} \frac{\partial J}{\partial \boldsymbol{H}} &= 2 \sum_{x=1}^{X'} \sum_{y=1}^{Y'} \boldsymbol{b}(x,y) \Big[\boldsymbol{b}(x,y)^{\mathrm{T}} \boldsymbol{H} - \boldsymbol{R}(x,y) \Big] \\ &+ \lambda \frac{\partial D(\boldsymbol{H})}{\partial \boldsymbol{H}}. \end{aligned} \quad (11)$$

III. Experiments

To investigate the effectiveness of the proposed method, we conducted two kinds of experiments. In the first experiment, we examined the quality of the super-resolution images using a Peak Signal-to-Noise Ratio (PSNR) criterion. On the other hand, the second experiment evaluated the recognition rates when the super-resolution images were input to a face recognition system.

A. Experimental schema

We took 870 facial videos of 29 subjects with a digital video camera while the face poses changed from $-20°$ to $20°$. These videos included slight changes of facial expressions. The face region size was approximately 64×64 pixels. Each video consisted of 30 frames. Figure 4 shows examples of the video frames used in the experiments.

As the input videos, we generated low-resolution videos by blurring and down-sampling each original video so that the face region sizes would be within 16×16 pixels and 32×32 pixels. Figure 5 shows an example of the low-resolution video frame.

Super-resolution images were generated so that the face region size would become 64×64 pixels. We manually selected a frame nearest to the frontal face as the reference image for the image alignment of the super-resolution for each input video. We compared the performance of the following three methods:

- Comparative method 1 (w/o SR): The low-resolution reference image was used as the super-resolution image.
- Comparative method 2 (w/o FFD): The image alignment was performed without the FFD method (only with the ICIA method).
- Proposed method (w/ FFD): The image alignment was performed with the FFD method.

We calculated the PSNR of the super-resolution image to the original high-resolution image of the reference image for each input video. On the other hand, in the face recognition experiment, the Eigenface method [11] was employed for the face recognition framework. We made an image set composed of 329 faces for the gallery images of face

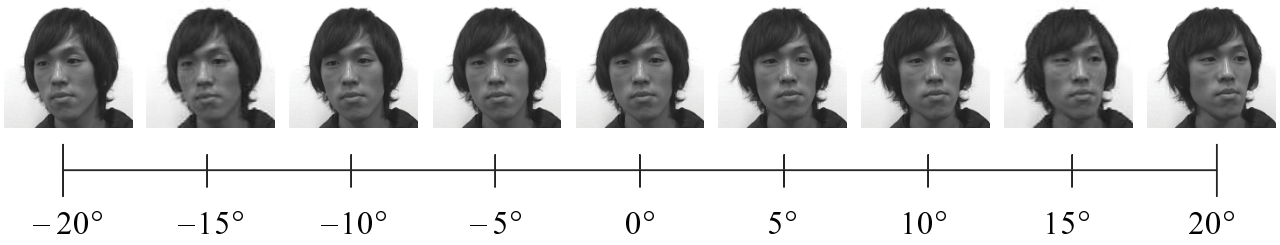

Figure 4. Examples of the video frames.

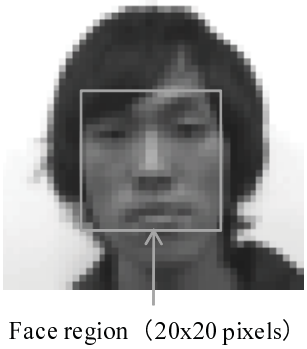

Figure 5. An example of a frame from input videos.

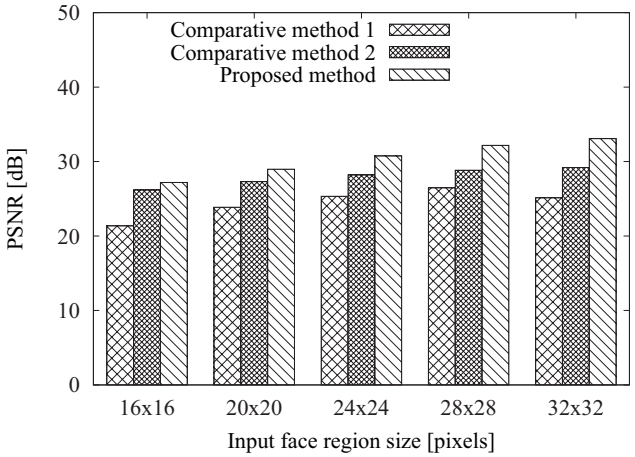

Figure 6. The PSNRs for each resolution.

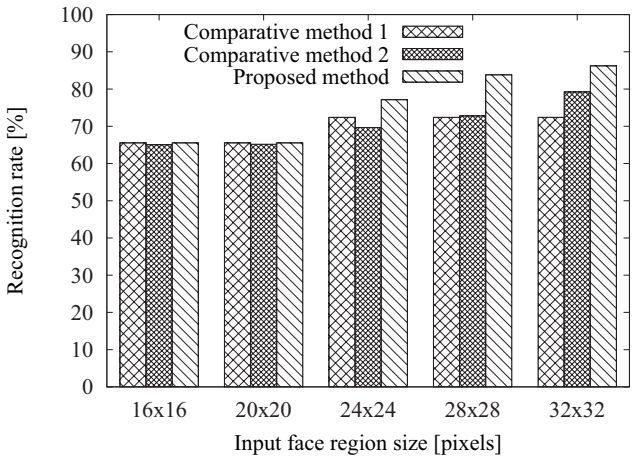

Figure 7. The recognition rates for each resolution.

recognition and the training images of the Eigenspace. The image set consisted of frontal face images of 29 subjects that were taken by a digital still camera different from the camera for the input videos, and 300 frontal face images from the face image dataset provided by SOFTPIA JAPAN.

B. Experimental results

Figure 6 shows the comparison of the PSNRs for each input face region size. For all face region sizes, the proposed method provided higher PSNRs than the comparative methods did. On the other hand, Figure 7 shows the comparison of the recognition rates for each resolution. When the input face region sizes were larger than 20×20 pixels, the recognition rates of the proposed method were higher than those of the comparative methods. Regarding the computation time, it took approximately 6 seconds to generate one super-resolution image by the proposed method.

IV. Discussion

A. Improvement of image qualities

From Figure 6, compared with the comparative method 1, the proposed method provided higher PSNRs, which improved 6.01dB on average. This indicates that the face image qualities were improved by the super-resolution. Compared with the comparative method 2, the proposed method improved the PSNRs 2.49dB on average. From this, we confirmed that the proposed method achieved accurate alignment between images including changes of face poses and expressions by using the FFD method. Figures 8 shows examples of the super-resolution images output from each method. In this figure, (b) and (c) demonstrate that the proposed method could generate less blurred images than the

(a) Comparative method 1 (w/o super-resolution) (b) Comparative method 2 (w/o FFD) (c) Proposed method (d) Original

Figure 8. Examples of the super-resolution images (Face region size: 32×32 pixels $\rightarrow$ 64×64 pixels).

comparative method 2 did, and could accurately reconstruct high-frequency components.

B. Improvement of recognition rates

When the input face region size for super-resolution was greater than 20×20 pixels, the proposed method provided higher recognition rates than the comparative methods did. Especially for face regions with a size of 32×32 pixels, the accuracy improved 13.8% compared with the comparative method 1 and 7.0% compared with the comparative method 2. This indicates the effectiveness of the proposed method for face recognition from low-resolution videos. However, significant improvement of the recognition rates of the proposed method for the comparative methods were not observed when the face region sizes were 20×20 pixels and 16×16 pixels. Figure 5 shows an example of a frame of the input video with a face region with a size of 20×20 pixels. In this case, we consider that sufficient information for accurate recognition could not be obtained from each frame of the input video because the change of the observed pixel values reduced in proportion to the decrease of the resolution even if changes of face poses and expressions occurred. Solutions for this problem would include super-resolution techniques based on learning [1]–[4]. These could use common characteristics of face images that are learnt from a large amount of face images. The performance of the proposed method would further be improved by using these techniques.

C. Cause of alignment failures

There were some cases that the alignments of face images failed due to a rapid change of expression, such as blinking, between the reference image and the target image. This causes a situation where a part of pixel correspondences is not available between the images. This situation decreases the alignment accuracy because the displacements of control points of the FFD method could not be calculated correctly. To solve this problem, we should make use of knowledge specific to face, such as positions and motions of facial parts. By adding constraints to the FFD method based on such knowledge, the ability of the alignment in such a situation would be enhanced.

V. Conclusion

We proposed a video-based super-resolution method for low-quality face images that can deal with non-rigid deformations caused by changes of face poses and expressions without any positional information of facial feature points. To achieve this, image alignment of each local region

between video frames is performed by using a Free-Form Deformation (FFD) method.

Experiments were conducted by using actual facial videos including the non-rigid deformations. The results of image quality evaluation demonstrated that the proposed method could provide high-quality super-resolution images. Moreover, from the results of face recognition experiments, we confirmed the effectiveness of the proposed method for face recognition from low-resolution videos.

Future work will include introduction of facial knowledge into the FFD method, such as positions and motions of facial parts, and combination of the proposed method with learning-based super-resolution techniques.

ACKNOWLEDGMENT

This work was supported by "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 face image dataset used in this work was provided by SOFTPIA JAPAN.

REFERENCES

[1] S. Baker and T. Kanade, "Hallucinating faces," in *Proceedings of 4th International Conference on Automatic Face and Gesture Recognition*, Mar. 2000, pp. 83–88.

[2] A. Chakrabarti, A. N. Rajagopalan, and R. Chellappa, "Super-resolution of face images using kernel PCA-based prior," *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 888–892, 2007.

[3] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes III, and R. M. Mersereau, "Eigenface-based super-resolution for face recognition," in *Proceedings of 2002 IEEE International Conference on Image Processing*, Sep. 2002, vol. 2, pp. 845–848.

[4] H. Huanga, H. Hea, X. Fanb, and J. Zhang, "Super-resolution of human face image using canonical correlation analysis," *Journal of Pattern Recognition*, vol. 43, no. 12, pp. 2532–2543, 2010.

[5] F. W. Wheeler, X. Liu, and P. H. Tu, "Multi-frame super-resolution for face recognition," in *Proceedings of First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, Sep. 2007, pp. 1–6.

[6] P. Mortazavian, J. Kittler, and W. Christmas, "3D-assisted facial texture super-resolution," *Signal Processing*, pp. 1–11, 2009.

[7] R. R. Schulz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Transactions on Image Processing*, vol. 5, no. 6, pp. 996–1011, 1996.

[8] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[9] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[10] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[11] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 16, pp. 71–86, 1991.