Super-resolution Restoration of Facial Images in Video

Jiangang Yu and Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, California 92521, USA
{bhanu,jyu}@vislab.ucr.edu

Abstract

Reconstruction-based super-resolution has been widely treated in computer vision. However, super-resolution of facial images has received very little attention. Since different parts of a face may have different motions in normal videos, this paper proposes a new method for enhancing the resolution of low-resolution facial image by handling the facial image non-uniformly. We divide low-resolution face image into different regions based on facial features and estimate motions of each of these regions using different motion models. Our experimental results show we can achieve better results than applying super-resolution on the whole face image uniformly.

1. Introduction

There is a growing interest in face recognition and identification for surveillance system, information security, and access control applications. Recognition of faces from video sequence utilizing both spatial and temporal information started only a few years ago [10]. There still exist difficult problems such as facial expression variations, occluded faces, different pose and lighting changes that need further investigation. Especially, the quality of video is usually low and face images are small, which raises the problem of enhancing low-resolution image from the video sequence.

Video sequences usually contain a large overlap between successive frames. Multiple captured images using motion of the sensors or objects potentially provide enough samples for super-resolution. However, super-resolution(SR) reconstruction is one of the most difficult and ill-posed problems due to the demand of accurate alignments between multiple images and multiple solutions for a given set of images. In particular, human face is much more complex compared to other objects which are addressed by the majority of the super-resolution literature. Super-resolution from facial images may suffer from subtle facial expression variation, non-rigid complex motion model, visibility and occlusion, and illumination and reflectance variations. Due to these reasons, most of existing super-resolution algorithms are not applicable to facial video sequences.

This paper is an attempt to tackle the problems brought by the complexity of facial images. We propose a method to treat facial image non-uniformly corresponding to different facial features. We use different motion models for different regions of facial image to align the facial images. Finally we perform super-resolution on the aligned images.

This paper is organized as follows. In section 2 we introduce related work and motivation. Also, we explain our contributions. In section 3, we describe our super-resolution algorithm and technical details. Experiments and analysis are presented in section 4. Section 5 provides the conclusion of this paper.

2. Related work, Motivation and Contributions

2.1 Related Work

Tsai and Huang [8] first addressed the problem of enhancing resolution for a sequence of low-resolution (LR) images of a translated scene. Since then there are numerous approaches focusing on this problem. Work on super-resolution (SR) can be categorized into two classes: reconstruction-based methods [8][5][6] and learning-based methods [1][4][7]. The former reconstructs the original signal based on sampling theory while the latter creates the signal based on learned model from samples [11]. In contrast with reconstruction-based methods, learning-based methods have two steps: training and testing. In training step, a generative model is learned. For testing step, the generative model is applied to LR images to acquire SR images. The major class of super-resolution algorithms is reconstruction-based methods. In this class there are frequency domain methods such as [8] and spatial domain paradigms such as [5][6].

2.2 Motivation

In the super-resolution literature, there are only a few approaches focusing on super-resolution of facial images. Baker and Kanade [1] propose learned-based super-resolution algorithm called hallucination or recogstruction

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

on human faces. Following this work, Dedeoglu et al. [4] use graphical model to encode spatial-temporal consistency of the low-resolution images and image formation & degradation processes. The nodes in the graphical model represent different patches of the images. The algorithm finds the best-matched patch in training set for the probe image. Park and Lee [7] propose a method of synthesizing high-resolution facial image using an error back-projection of example-based learning. The above super-resolution methods of facial images are all learning-based methods. They need a certain amount of training faces and assume alignment is done before they apply their methods. However, accurate alignment is the most critical step for super-resolution techniques. Since features of facial images from video sequence may undergo different complex motion (this is usually true), it is necessary that we handle super-resolution of facial image non-uniformly corresponding to different facial regions. Figure 1 shows five facial frames from one video sequence with corresponding low-resolution frames[1]. It is clear that the facial features such as eyes, eyebrows and mouth undergo different motions in these frames.

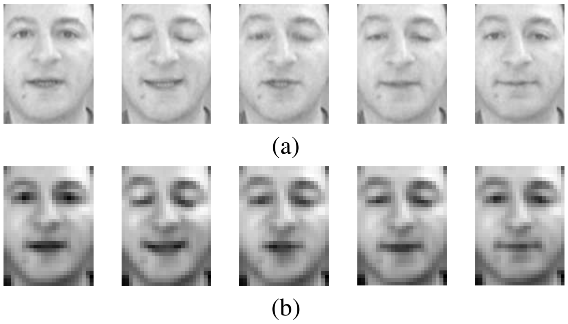

Figure 1. high-resolution images with corresponding low-resolution images. (a) high-resolution images. (b) low-resolution images.

2.3 Contributions

In normal video, human face may undergo different motions for different parts of a face. As global super-resolution approaches, they do not address the problem of how to select appropriate facial image regions in which to apply appropriate motion models and implement super-resolution. In this paper, we treat human face non-uniformly corresponding to facial features. We segment face based on facial features and align them separately. Due to the non-rigidity of human face, we can better interpret the motions of the facial images than global approaches with the end result of having a better high resolution face image.

3. Technical Approach

Figure 2 shows the block diagram of our approach. As shown in figure 1, facial image has different motions for different facial regions due to non-rigidity of human face. We first divide facial image into six patches including left/right eyebrow, left/right eye, mouth and the other part of the face. Following this step we track different patches in frames and align them separately using different motion models. Subsequently we reconstruct the super-resolution image on the aligned facial images.

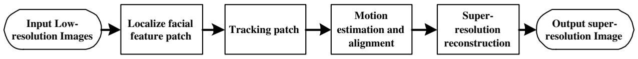
Input Low-resolution Images
Localize facial feature patch
Tracking patch
Motion estimation and alignment
Super-resolution reconstruction
Output super-resolution Image

Figure 2. Block diagram of our approach.

3.1 Facial regions tracking

In our work, a plane tracking algorithm based on minimizing the sum of squared difference between stored image of the patch and the current image of it [3] is used for tacking different facial patches. The motion parameters are solved by minimizing the sum of squared difference between the template and the rectified image. Through Jacobian Matrix factorization in [3] we can projectively track a planar patch from the consecutive frames using projective motion model. In figure 3, we show the result of tracking the left/right eyebrow, left/right eye and the mouth of the facial images.

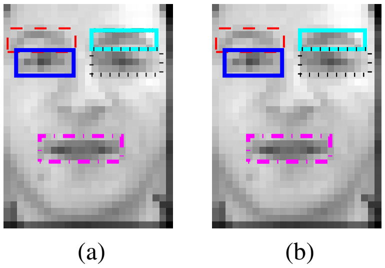

Figure 3. Tracking of eyes, eyebrows and mouth in two frames. (a) first frame. (b) second frame.

3.2 Motion Estimation

The correspondence maps between the low-resolution patches are established using optic flow techniques. The tracked patches which correspond to the same regions are warped towards the desired low-resolution image. Black et al. [2] propose one of the sophisticated optical flow techniques called robust optic flow. It employs statistics to avoid

[1] The data is recorded by authors in [4]

large errors caused by outliers and allows intensity discontinuities in the assumption of flow field computation. Considering the non-rigidity of human face, we employ different motion models for different parts of facial images. For eyebrows, eyes and mouth, we employ projective motion model. As for the other part of the face, we use affine model to estimate the motion between them.

Even though a robust flow techniques is used, there may exist warping errors due to the violation of basic assumptions such as Lambertian surface, particularly for low-resolution images. For facial images, people may have different facial expressions in consecutive frames or even worse situation such as closing eyes. To some extent that we partition facial image into different facial feature regions can relieve this aspect of error. In order to detect anomalies in flow based warping, we design a match statistics to measure how well the warped patches align with the target patches. We compute the normalized mean square error between a target patch and the warped patch. If the matching score at a patch is below a certain threshold (we use 0.90), the warped patch indicates misalignment and will be ignored in the super-resolution process.

3.3 Iterative Back-projection (IBP) reconstruction

Our super-resolution algorithm is based on IBP [7]. The input images to IBP are the motion compensated facial image introduced in section 3.2. The iterative update of super-resolution image is defined in [7] as:

$$\mathbf{X}^{(n+1)} = \mathbf{X}^{n} + \frac{1}{K}\sum_{k=1}^{K}(((y_k - y_k^{(n)}) \uparrow \mathbf{s}) * \mathbf{p})^{\mathbf{F}_k} \quad (1)$$

where $\mathbf{X_n}$ denotes the recovered high-resolution image at the nth iteration, y_k is the input low-resolution image, K is the number of low-resolution images, $\uparrow \mathbf{s}$ represents an up-sampling operator by a factor $\mathbf{s}, \mathbf{p}$ is a back-projection kernel, and $\mathbf{F}_k$ is forward-warping process. The simulated image $\mathbf{X}^n$ is generated as follows:

$$\mathbf{Y}_n = ([\mathbf{X}_n]^{\mathbf{B}_k} * \mathbf{h}) \downarrow s \quad (2)$$

where $\mathbf{B}_k$ is a backward warping process, $\mathbf{h}$ is a blurring kernel, and $*$ is the convolution operator. The super-resolution process is repeated iteratively to minimize the error function:

$$\mathbf{E}^n = \sqrt[2]{\frac{1}{K}\sum_{k=1}^{K}||Y_k - Y_k^{(n)}||_2^2} \quad (3)$$

4. Experimental Results

We implement our experiments with two types of datasets: semi-synthetic data and real data.

- Semi-synthetic data has synthetically sub-sampled images but real motion. This is generated from a real sequence of high-resolution images.

- Real data are low-resolution images taken by a low-end consumer camera.

4.1 Semi-synthetic data

We use the dataset in [4], which is a 40 seconds video sequence of a speaking person. We use 26 frames as input LR images. The face in the video sequence covers an area of 128x96 pixels. Figure 1 shows some examples of this video sequence. We blur and down-sample the video sequence at a resolution of 32x24 pixels. In the first frame, we divide the face into six regions and track them in the consecutive frames using the method introduced in section 3.1. Then we use robust motion estimation method to align the six regions with respect to the reference frame. IBP super-resolution algorithm is subsequently implemented to acquire the high-resolution image.

Figure 4 shows one of the low-resolution images (a), Bicubic interpolated high-resolution image (b), reconstructed high-resolution image uniformly (c) and the high-resolution image (d) acquired by our method. Compared with (d), the upper eyelids of high-resolution image in (c) occludes the eyeballs of the person. For the person's lips, (c) is much blurred than that in (d). The reason why this happens in (c) is due to the different motions for different parts of facial images. The person may close his eyes or mouth in course of video recording, which results in differences in consecutive facial frames, especially for eyes, eyebrows and mouth. If we simply discard the frames in which the person's eyes or mouth are closed, this is not a wise choice because the rigid part of the face can still contribute to the super-resolution reconstruction. This is one of the reasons why we design our algorithm to handle human face non-uniformly.

In order to measure the performance of our algorithm, we compute peak signal-to-noise ratio (PSNR) and Structural SIMilarity (SSIM) index [9] as measurements between target HR and reconstructed SR images. The simplest and most widely used quality metrics are mean squared error (MSE) and PSNR. Mean squared Error (MSE) and PSNR are simple to compute and have clear physical meanings. However, they are not very well matched to perceived visual quality. SSIM is a method that provides a quality measurement of images based on their structural contents. Since the structures of the objects in the scene are independent of the illumination, SSIM seperates the influence of the illumination. As SSIM is based on structural content rather than Mean Squared Error with error weighted by different visibility models of the human visual system, it does not suffer from this issue, yet is strongly correlated with per-

Table 1. Performance Comparison.

Methods	Quality Measurements	
	PSNR	**MSSIM**
Bi-cubic interpolation	17.658	0.77015
Uniform reconstruction	14.571	0.74585
Non-uniform reconstruction	26.742	0.87583

ceptual image quality. Table 1 represents the measurements in terms of PNSR and SSIM. From Table 1, our proposed method has better score than the other two techniques in terms of PSNR and SSIM, which is conformed also by the visualization of Figure 4. For SSIM, the highest score happens when two images are identical.

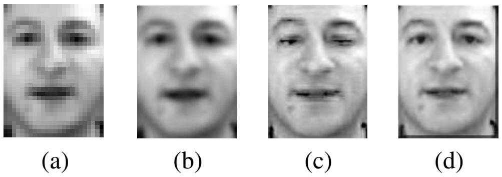

Figure 4. Comparison of reconstruction results on semi-synthetic data. (a) Low-resolution image. (b) Bi-cubic interpolated image. (c) Uniformly reconstructed image. (d) Non-uniformly reconstructed image.

4.2 Real Data

The real data is obtained by a low-end consumer camcorder. The person is walking towards the camera while talking. The distance between the camera and the walking person is about 30 feet. Figure 5 shows one low-resolution image, uniformly reconstructed HR image and HR image of our method.

5. Conclusions

In this paper, we propose a method of enhancing the resolution of low-resolution facial images through handling the facial image non-uniformly. We segment facial image into different regions corresponding to different motion models and estimate the motions non-uniformly of tracked regions in the consecutive frames. The experimental results provide a proof of the concept for our method and show that our method gives better results than handling the face uniformly. For low-quality real data, how to automatically segment facial image and track the features is an open problem for us. In the future, we will develop automatic methods for accomplishing it.

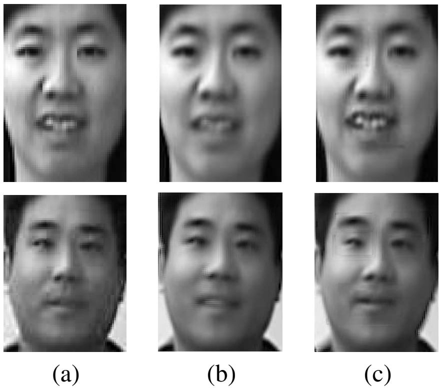

Figure 5. Comparison of reconstruction results on real data. (a) Low-resolution image. (b) Uniformly reconstructed image. (c) Non-uniformly reconstructed image.

References

[1] S. Baker and T. kanade. Limits on super-resolution and how to break them. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24:1167–1183, 2002.

[2] M. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[3] J. Buenaposada and L. Baumela. Real-time tracking and estimation of plane pose. *Proc. Int. Conf. on Pattern Recognition*, 2:697–700, 2002.

[4] G. Dedeoglu, T. Kanade, and J. August. High-zoom video hallucination by exploiting spatio-temporal regularities. *7th IEEE Workshops on Application of Computer Vision*, 2:151–158, Jun. 2004.

[5] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy and undersampled measured images. *IEEE Trans. on Image Processing*, 6(12):1646–1658, Dec. 1997.

[6] M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation*, 4:324–335, 1993.

[7] J. Park and S. Lee. Resolution enhancement of facial image using an error back-projection of example-based learning. *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 831–836, 2004.

[8] R. Tsai and T. Huang. Multiframe image restoration and registration. *Advances in Computer Vision and Image Processing*, 1:317–339, 1984.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on Image Processing*, 27(4):619–624, 2005.

[10] W. Zhao, R. Chellapa, and P. Phillips. Face recognition: A literature survey. *ACM Computing Survey*, 35(4):399–458, Dec. 2003.

[11] W. Zhao and H. S. Sawhney. Is super-resolution with optical flow feasible? *Proc. of the 7th European Conf. on Computer Vision*, 1:599–613, May 2002.