SUPER-RESOLVED FACIAL TEXTURE UNDER CHANGING POSE AND ILLUMINATION

Jiangang Yu, Bir Bhanu, Yilei Xu and Amit K. Roy-Chowdhury

Center for Research in Intelligent System, University of California, Riverside, CA 92521

ABSTRACT

In this paper, we propose a method to incrementally super-resolve 3D facial texture by integrating information frame by frame from a video captured under changing poses and illuminations. First, we recover illumination, 3D motion and shape parameters from our tracking algorithm. This information is then used to super-resolve 3D texture using Iterative Back-Projection (IBP) method. Finally, the super-resolved texture is fed back to the tracking part to improve the estimation of illumination and motion parameters. This closed-loop process continues to refine the texture as new frames come in. We also propose a local-region based scheme to handle non-rigidity of the human face. Experiments demonstrate that our framework not only incrementally super-resolves facial images, but recovers the detailed expression changes in high quality.

Index Terms— Super-resolution, 3D, Facial image

1. INTRODUCTION

Face recognition and identification for surveillance systems, information security, and access control has received growing attention. In many of the above scenarios, the distance between the objects and the cameras is quite large, which makes the quality of video usually low and facial images small. Zhao et al. [1] identify low-resolution (LR) as one of the challenges in video-based face recognition. Super-resolution (SR) from multiple images in video has been studied by many researchers in the past decades. There still exist problems such as facial expression variations, different poses and lighting changes that need further investigation. In our approach, we propose a closed-loop framework that super-resolves facial texture through the combined effects of motion, illumination, 3D structure and albedo.

Our method super-resolves facial texture utilizing both spatial and temporal information from video through changing poses and illuminations. We track the image sequence and estimate illumination parameters through 3D tracking, interpolation by changing pose and illumination normalized images. We also use local-region based super-resolution to handle the non-rigidity of human face.

2. RELATED WORK, MOTIVATION AND CONTRIBUTIONS

2.1. Related work

Based on spatial or frequency domains, we categorize SR approaches into two classes: spatial domain and frequency domain. SR approaches can also be divided into reconstruction-based and learning-based methods based on whether training step is employed. Schultz and Stevenson [2] use a Huber-Markov-Gibbs model for the $a\ priori$ model to preserve edges while achieving a global smoothness constraint. Another approach toward the SR reconstruction problem is the method of projections onto convex sets (POCS) [3]. Irani and Peleg [4] propose an iterative back-projection (IBP) method updating the estimate of the SR reconstruction by back-projecting the error between the simulated LR images and the observed LR ones. Yu and Bhanu [5] adopt this method for super-resolving 2D facial images non-uniformly based on the local regions. There exist hybrid methods which combine ML/MAP/POCS based approaches to SR reconstruction [2]. In the SR literature, there are only a few approaches that focus on super-resolution of facial images. Baker and Kanade [6] propose learning-based SR algorithm named hallucination or *recogstruction* on human facial images. Following this work, Dedeoglu et al. [7] adopt graphical model to encode spatial-temporal consistency of the LR images. The above methods are all learning-based SR approaches and need a certain amount of training faces. They assume alignment is done before applying SR methods. However, accurate alignment is the most critical step for SR techniques.

2.2. Motivation and Contribution

We propose a framework to incrementally super-resolve facial video under changing illumination and pose. Unlike traditional approaches which extract SR frames from multiple images using a "sliding window" with respect to a reference frame, we integrate spatial and temporal information of LR frames to refine 3D facial texture for the entire video.

LR video is usually taken under uncontrolled condition at a distance. Hence there may be large illumination and pose variation in the acquired video. Traditional motion estimation techniques in the existing SR literature that use dense flow or parametric transformation without compensating for illumination changes will not work at the registration stage. And

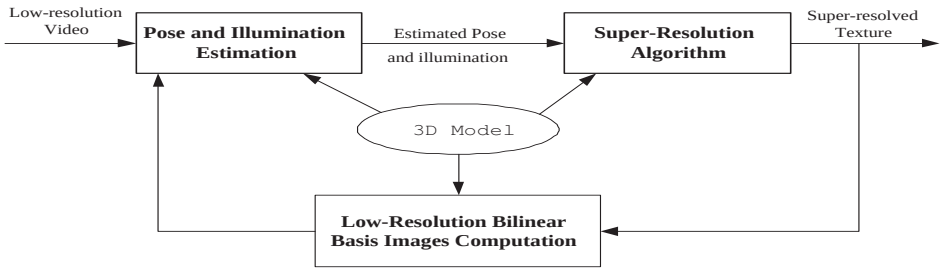

Fig. 1. Block diagram of our approach.

it is difficult to handle large pose changes in video for these techniques. We estimate 3D motion and illumination parameters for the video to register the images with a generic 3D model and normalize for the illumination. We also design a scheme to take special care of the non-rigidity of human face with expression changes.

3. TECHNICAL APPROACH

The block diagram of our approach is shown in Figure 1. A generic 3D face model [8] is used in our approach. This generic model is acceptable for super-resolving the 3D texture. The problem of obtaining a more accurate 3D structural model is not the focus of this paper. We first track the pose and estimate illumination of the incoming frame from a video. Then the tracked pose and estimated illumination are passed to the super-resolution algorithm for super-resolving the 3D facial texture. Following this step the super-resolved 3D facial texture is fed back to generate low-resolution bilinear basis images, which are used for pose tracking and illumination estimation. The feedback process improves the estimates of pose and illumination in subsequent frames. This process is continuously repeated to refine the 3D facial texture as new frames come in. Note that our tracking and super-resolution algorithms are 3D based approaches.

3.1. Bilinear Basis Images Computation

It has been proved that for a fixed Lambertian object, the set of reflectance images *under distant lighting without cast shadow* can be approximated by a linear combination of the first nine spherical harmonics [9]. In recent work [10], motion was taken into the consideration in the above formulation. It was shown that for moving objects it is possible to approximate the sequence of images by a bilinear subspace using tensor notation as

$$\mathcal{I} = \left(\mathcal{B} + \mathcal{C} \times_2 \begin{pmatrix} \mathbf{T} \\ \mathbf{\Omega} \end{pmatrix}\right) \times_1 \mathbf{l}, \tag{1}$$

where $\times_n$ is the *mode-n product* . $\mathcal{I} \in \mathbb{R}^{1\times 1\times M\times N}$ is a sub-tensor representing the image, $\mathcal{B} \in \mathbb{R}^{N_l\times 1\times M\times N}$ is a sub-tensor comprising the illumination basis images. $\mathcal{C} \in \mathbb{R}^{N_l\times 6\times M\times N}$ incorporates the bilinear basis for the motion and illumination, and $\mathbf{l} \in \mathbb{R}^{\mathbf{9}}$ is the vector of illumination coefficients.

Thus, low-resolution bilinear basis images are obtained from the super-resolved texture and passed on to the pose and illumination component described below.

3.2. Pose and Illumination Estimation

The joint illumination and motion space described above provides us with a method for estimating 3D motion of moving objects in video sequences under time-varying illumination conditions as:

$$\begin{aligned}\left(\hat{\mathbf{l}}, \hat{\mathbf{T}}, \hat{\mathbf{\Omega}}\right) &= \arg\min_{\mathbf{l},\mathbf{T},\mathbf{\Omega}} \|\mathcal{I}_{t_2} - \left(\mathcal{B}_{t_1} + \mathcal{C}_{t_1} \times_2 \begin{pmatrix} \mathbf{T} \\ \mathbf{\Omega} \end{pmatrix}\right) \times_1 \mathbf{l}\|^2 \\ &\quad +\alpha\| \begin{pmatrix} \mathbf{T} \\ \mathbf{\Omega} \end{pmatrix} \|^2 \end{aligned} \tag{2}$$

After this process, 3D motion $\hat{\mathbf{T}}$ and $\hat{\mathbf{\Omega}}$ along with illumination coefficients $\mathbf{l}$ are estimated. We use the estimated motion and illumination to get the illumination normalized frame with respect to the reference illumination and pass it along with the 3D motion estimate to the super-resolution algorithm.

3.3. Super-resolution Algorithm

We adopt IBP [4] algorithm and extend it to 3D as our SR method. Due to the non-rigidity of the face, we reconstruct SR texture separately based on six facial regions of the face. The inputs to IBP are illumination normalized LR images and the current super-resolved texture. The block diagram is illustrated in Figure 2.

In Figure 2, $\mathbf{X}_n$ is the currently reconstructed 3D facial texture at the n-th frame. We define $\mathbf{Y}_n^k$ as the k-th illumination normalized LR facial region. $\hat{\mathbf{T}}$ and $\hat{\mathbf{\Omega}}$ represent the tracked pose passed from the tracking algorithm. $\mathbf{B}_n$ is the process of projecting 3D texture to form 2D image while $\mathbf{B}_n^{-1}$ denotes the inverse process. $\mathbf{h}$ is the blurring function and $\mathbf{P}$ denotes the back-projection kernel as proposed in [4]. $\uparrow\mathbf{s}$ represents an up-sampling operator by a factor $\mathbf{s}$.

Due to the non-rigidity of a human face, there may exist facial expression changes such as closing eyes and opening

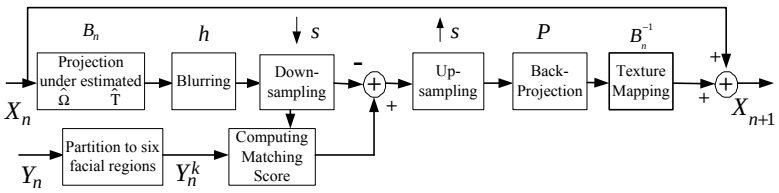

Fig. 2. Block diagram of the super-resolution algorithm.

mouth. In order to handle facial expression changes in images, we use local-based SR approach by dividing facial image into different regions (two eyes, two eye brows, mouth and the rest of the face) based on facial features. We interactively locate eyes, eye brows, and mouth in the 3D model during registration of the *first frame* to 3D model. For each incoming LR image, we calculate a match statistic to detect whether there are significant expression changes. If the match score is below a certain threshold, the corresponding part will be ignored during super-resolving the texture. According to [4], the revising values of model texture $\Delta \mathbf{X}_n$ are calculated by simulated pixel values from 3D texture and the observed image pixels (illumination normalized) as given by the following equation:

$$\Delta \mathbf{X}_n = \bigcup_{k=1}^{6} (((\mathbf{Y}_n^k - \tilde{\mathbf{Y}}_n^k) \uparrow s) * \mathbf{P})^{\mathbf{B}_n^{-1}} \tag{3}$$

The simulated image$\tilde{\mathbf{Y}}_{n'}^k$ is generated as:

$$\tilde{\mathbf{Y}}_n^k = ([\mathbf{X}_n^k]^{\mathbf{B}_n} * \mathbf{h}) \downarrow s \tag{4}$$

We define our match measure as follows,

$$\mathbf{E}_k = \sum_{x=1}^{M} \sum_{y=1}^{N} \frac{\left((\mathbf{Y}_n^k(x,y) - \mu_1)([\mathbf{X}_n^k(x,y)]^{\mathbf{B}_n} - \mu_2)\right)}{M{*}N{*}\sigma_1{*}\sigma_2} \tag{5}$$

where M and N are the image size, μ_1 and μ_2 are respective means of image regions, σ_1 and σ_2 are respective image variances within the region.

4. EXPERIMENTAL RESULTS

We carry out our a variety of experiments to demonstrate the performance of our closed-loop approach.

4.1. Synthetic Data

Given pose changes, we generate LR video sequence synthetically from scanned 3D model [8] with high-resolution texture. The generated LR video sequence is blurred by a 7x7 Gaussian blur function and down-sampled to about 30x30 pixels.

Not only does our approach integrate information from commonly visible part on the images, it can also integrate invisible texture on the previous pose to super-resolve the 3D texture. The last row in Figure 3 shows this tendency from the first reconstructed image to the last one. On the first reconstructed SR image, the left-most part is black and blurred because this part is not visible from the previous input LR images. After some iterations, it is getting better and better as new frames at visible poses are available. Figure 5 clearly shows this tendency through the measurement of peak signal-to-noise ration (PSNR) between reconstructed SR image with target SR ones. PSNR of common face parts denotes the frontal facial region for the face. PSNR of non-common parts represents the facial region which becomes completely visible at the end from being invisible in the first frame. From this figure, PSNR of the common parts keeps rising higher after the first 41 frames and almost keep constant at a value close to 31. Compared with the initial guess of the super-resolved texture, the one after the first 41 frames integrate information from these frames which causes PSNR value to go higher. PSNR of common face parts goes higher again at about 100-th frame because the occluded region of common face parts becomes visible. PSNR of non-common parts shows the process of super-resolving for the invisible part of the face in the first frame.

4.2. Real video

We test our algorithm on a video of person whose face has significant expression over time. Assuming the face is a rigid object, we track this sequence over time during tracking. In super-resolution step, we use a local-based approach to super-resolve the texture. The results are shown in Figure 4. We interactively locate the six regions in the 3D model during registration of the first frame as described in section 3.3. We then compute the match statistic between partitioned region of illumination normalized input LR image and corresponding super-resolved texture for the coming frames. If there are less than 10 continuous frames which are determined by the match statistic to have significant facial expression changes, we will discard the corresponding parts of LR image for super-resolution. Otherwise, we believe that they are valid expressions and we super-resolve the associated texture. In this situation, we refresh the previously constructed texture.

5. CONCLUSIONS

In this paper, we propose a video-based super-resolution approach where pose and illumination invariant tracking and super-resolution take place in a closed-loop. Our experimental results show that our method can acquire super-resolution video with this novel closed-loop system. Moreover, our method can handle the non-rigidity of human face since the facial im-

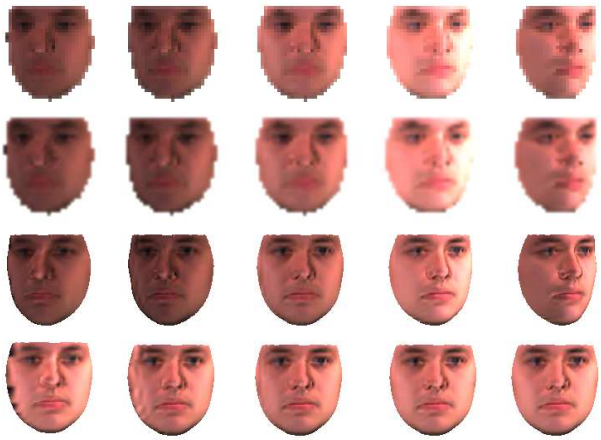

Fig. 3. Results on Synthetic video with Ground Truth Poses. The first row shows the original LR frames and the second row shows the bicubic interpolated ones. Reconstructed SR images are shown in the third row. The last row shows pose and illumination normalized reconstructed SR images with respect to the middle input LR image.

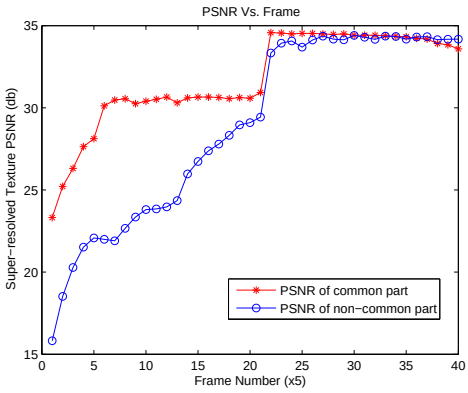

Fig. 5. Evaluation of super-resolved video as measured by the peak signal-to-noise ratio for the results shown in Fig. 3.

ages are processed non-uniformly for different regions of the face.

6. ACKNOWLEDGEMENT

This work is partially supported by NSF award CNS-0551741. The contents of the information do not reflect the position or policy of the U.S. Government.

7. REFERENCES

[1] W. Zhao, R. Chellapa, and P.J. Phillips, "Face recognition: A literature survey," *ACM Computing Survey*, vol. 35, no. 4, pp. 399–458, December 2003.

[2] R. R. Schultz and R. L. Stevenson, "Extraction of high resolution frames from video sequences," *IEEE Trans. Image Processing*, vol. 5, no. 6, pp. 996–1011, June 1996.

[3] M. I. Sezan A. J. Patti and A. M. Tekalp, "Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time," *IEEE Trans. Image Processing*, vol. 6, pp. 1064–1076, August 1997.

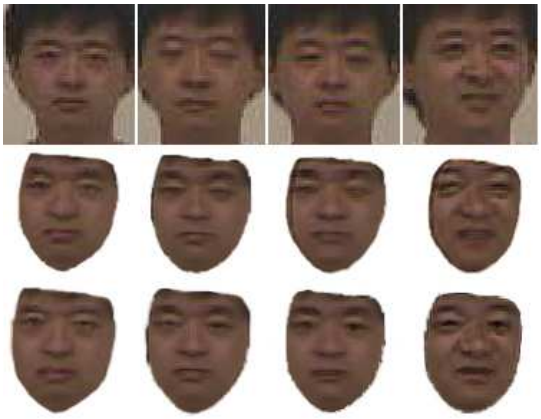

Fig. 4. Real video with expression changes. The first row shows the original LR frames and the second row shows the reconstructed ones with the global method. The third row shows the SR images of our local-based method.

[4] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *Journal Visual Communication Image Represent*, vol. 4, pp. 324–335, December 1993.

[5] J. Yu and Bir Bhanu, "Super-resolution restoration of facial images in video," *ICPR'06*, vol. 4, pp. 342–345, February 2006.

[6] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. PAMI*, vol. 24, no. 9, pp. 1167–1183, September 2002.

[7] G. Dedeoglu, T. Kanade, and J. August, "High-zoom video hallucination by exploiting spatio-temporal regularities," *IEEE CVPR'04*, vol. 2, pp. 151–158, Jun. 2004.

[8] V. Blanc and T. Vetter, "A morphable model for the synthesis of 3d faces," *Computer Graphics Proc. SIGGRAPH '99*, pp. 187–194, 1999.

[9] R. Basri and D.W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. PAMI*, vol. 25, no. 2, pp. 218–233, February 2003.

[10] Y. Xu and A. Roy-Chowdhury, "Integrating the effects of motion, illumination and structure in video sequences," *ICCV*, 2005.