SUPER-RESOLUTION OF DEFORMED FACIAL IMAGES IN VIDEO

Jiangang Yu and Bir Bhanu

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

ABSTRACT

Super-resolution (SR) of facial images from video suffers from facial expression changes. Most of the existing SR algorithms for facial images make an unrealistic assumption that the “perfect” registration has been done prior to the SR process. However, the registration is a challenging task for SR with expression changes. This paper proposes a new method for enhancing the resolution of low-resolution (LR) facial image by handling the facial image in a non-rigid manner. It consists of global tracking, local alignment for precise registration and SR algorithms. A B-spline based Resolution Aware Incremental Free Form Deformation (RAIFFD) model is used to recover a dense local non-rigid flow field. In this scheme, low-resolution image model is explicitly embedded in the optimization function formulation to simulate the formation of low resolution image. The results achieved by the proposed approach are significantly better as compared to the SR approaches applied on the whole face image without considering local deformations.

***Index Terms*—** Super-resolution, Free-form deformation

1. INTRODUCTION

Super-resolution reconstruction is one of the 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 have been used in the majority of the super-resolution literature. Super-resolution from facial video may suffer from subtle facial expression variation, occlusion, illumination and reflectance variations. Figure 1 shows six low-resolution facial frames from one video sequence with the corresponding high-resolution frames. It is clear that the face undergoes non-rigid motions.

In order to tackle the problems brought by the complexity of facial images, in this paper, we propose a novel global-to-local approach to locally align the images before applying SR algorithms. The approach consists of three steps: global tracking, local alignment and SR algorithm. In the global tracking step, a global transformation is used to track the face through the video sequence. Following the global registration, a space warping technique − Free Form Deformations (FFD) is used for modeling the local deformation of faces. The globally aligned image is parameterized using a set of B-spline functions and a set of control points overlaid on its volumetric embedding space. Correspondence is obtained through evolving a control lattice overlaid on the source image. We explicitly embed the LR image formation into the FFD formulation to simulate the process of LR imaging. We use three SR algorithms [1][2][3] in the last step and compare performance results on many real video sequences.

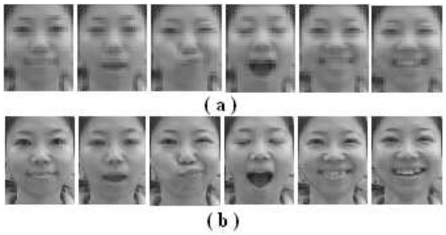

Fig. 1. An example of facial images with expression changes. (a) Low-resolution (30x24) images. (b) High-resolution (92x73) images.

1.1. Related Work and Contributions

In the past decades, a number of SR techniques have been proposed [3][4][5][6]. Based on whether a training step is employed in SR restoration, they are categorized as: reconstruction-based methods [1][3] and learning-based methods [4][5][6][7]. The major class of super-resolution algorithms is reconstruction-based methods. Markov Random Field (MRF) model is usually assumed for high resolution images in a learning-based method [5][7][8]. Most of the SR approaches focusing on face images are learning-based methods. The learning-based SR approaches need a certain amount of training data of faces. They do not handle local deformations. They assume that alignment has been performed before applying SR methods. However, accurate alignment is the most critical step for SR techniques.

Our proposed approach integrates the alignment and super-resolution steps. It combines a global parametric transformation with local deformation to cope with this problem. The key contributions of this paper are: (a) A hierarchical registration scheme that combines a global parametric transformation with a local free-form deformation (FFD).

978-1-4244-1764-3/08/$25.00 ©2008 IEEE

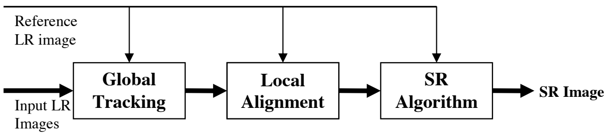

Fig. 2. Block diagram of our approach.

The global transformation tracks the face through the video sequence using a global motion model. A B-spline based Free-form Deformation is used to locally warp the input LR images to register with the reference LR image. (b) We explicitly embed low resolution imaging model in the formulation of FFD to simulate the formation of LR images. (c) Results are shown on a large number of videos.

2. TECHNICAL APPROACH

The overall approach is shown in Figure 2. Given a sequence of facial images, we first track the facial region using a global motion model [9]. After this global tracking step, we find the optimal transformation $\mathbf{T} : (x, y) \longrightarrow (x_0, y_0)$ which maps any point of the facial region in the image sequence $\mathbf{I}(x, y, t)$ at time t into its corresponding point in the reference image $\mathbf{I}(x_0, y_0, t_0)$. Reference image is the first image in the video sequence. We extract the facial region from the video sequence after the global tracking. Following this step, we deform the globally aligned facial regions (rectified image) to locally register with the regions in the reference image using our specially designed FFD algorithm that accounts the nature of LR data. The last step is to super-resolve the facial image on these registered images to acquire the SR image. Therefore, we design a combined transformation $\mathbf{T}$ consisting of a global transformation and a local deformation as follows

$$\mathbf{T}(x, y, t) = \mathbf{T}_{global}(x, y, t) + \mathbf{T}_{local}(x, y, t) \qquad (1)$$

2.1. Local Deformation

We consider an Incremental Free Form Deformation (IFFD) [10] formulation to model the local deformation and integrate it into our framework. In order to cope with the large local motion such as open mouth in Figure 1, we adopt a multi-level IFFD from coarse to fine control points levels. Given an incoming frame which is globally aligned, a control lattice P is overlaid on the image space. By evolving the control points in the lattice, the displacements of all the control points are acquired. Subsequently, B-spline basis functions are used as interpolation functions to get the dense deformation field.

Cost Function: Given the local deformation formulations, we need to find the deformation parameters of the control lattice $\Delta\mathbf{P}$. Then we can warp the input frame $\mathbf{I}(x, y)$ to register with the reference frame $\mathbf{I}(x_0, y_0)$ using the cubic B-spline functions. Similar to [10], we use Sum of Squared Differences (SSD) as the data-driven term for the optimization of energy function,

$$E_{data}(\Delta\mathbf{P}) = \iint_{\Omega} (\mathbf{I}(x, y) - g(\mathbf{T}(x, y; \Delta\mathbf{P})))^2 dxdy \qquad (2)$$

In order to account for outliers and noise, we consider an additional smoothness term on the deformation field δP as

$$E_{smoothness}(\Delta\mathbf{P}) = \iint_{\Omega} \left(\left\| \frac{\partial \delta \mathbf{T}}{\partial x} \right\|^2 + \left\| \frac{\partial \delta \mathbf{T}}{\partial y} \right\|^2 \right) dxdy \quad (3)$$

Combining (2) with (3), we can write the engergy term as

$$E(\Delta\mathbf{P}) = E_{data}(\Delta\mathbf{P}) + \lambda E_{smoothness}(\Delta\mathbf{P}) \qquad (4)$$

where λ is a constant which defines the tradeoff between the displacements and the smoothness of the transformation. The calculus of variations and a gradient descent method can be used to optimize the energy function. We can take the derivative of $E(\Delta\mathbf{P})$ with respect to the deformation parameters $\Delta\mathbf{P}$ as $\frac{\alpha E(\Delta\mathbf{P})}{\alpha \Delta\mathbf{P}^{[x]}_{(m,n)}}$ and $\frac{\alpha E(\Delta\mathbf{P})}{\alpha \Delta\mathbf{P}^{[y]}_{(m,n)}}$ to find the deformation $\Delta\mathbf{P}$ by minimizing the energy function.

Resolution Aware Local Deformation: In order to account for the complexity brought by LR data, we integrate LR image formation model into the FFD formulation. Considering the LR imaging model of a digital camera, LR images are blurred and sub-sampled (aliased) from the high resolution data with additive noise. While FFD works well with high resolution data, its accracy of deformation degrades quickly at low resolution. We integrate LR imaging model into the FFD formulation. We perform local deformations based on the following considerations:

Deform local motion on high resolution data: Without the loss of generality, we assume the camera is fixed and the relative motion between the object and the camera is due to the motion of the object. Local motion occurs on the object while the acquired LR image is the process of digital camera. To better model the local motion, instead of deforming the control lattice on LR image we perform FFD on the high resolution data. Then we simulate the process of LR imaging from the deformed high resolution image to get the motion compensated LR image. Plugging the LR imaging model into the data driven term in Equation (2), we can rewrite (2) as

$$\begin{aligned} & E_{data}(\Delta\mathbf{P}) \\ = & \iint_{\Omega} \left(\mathbf{I}_{LR}(x, y) - f(X, Y; \Delta\mathbf{P})\right)^2 dxdy \end{aligned} \qquad (5)$$

where

$$\begin{aligned} & f(X, Y; \Delta\mathbf{P}) \\ = & \quad g(\mathbf{T}(X, Y; \Delta\mathbf{P})) * \mathbf{h}) \downarrow_s \\ = & \iint_{(X',Y') \in bin(X,Y)} \varphi(\cdot) dX'dY' \end{aligned} \qquad (6)$$

In (6), $\varphi(\cdot)$ is defined as

$$\varphi(\cdot) = g(\mathbf{T}(X', Y'; \Delta\mathbf{P}))\mathbf{h}(X - X', Y - Y')dX'dY' \quad (7)$$

where (X, Y) is the pixel coordinates on high resolution image, $bin(X, Y)$ is the sensing area of the discrete pixel (X, Y) and $\mathbf{h}$ is the blurring function (Point Spread Function). The continuous integral in (6) is defined over $bin(X, Y)$ to simulate formation of LR image. The smoothness term in (3) is rewritten as

$$E_{smoothness}(\Delta\mathbf{P}) = \iint\limits_{\Omega} \phi(\cdot)dXdY \quad (8)$$

where

$$\phi(\cdot) \;=\; \left\| \frac{\partial \delta\mathbf{T}(X, Y; \Delta\mathbf{P})}{\partial X} \right\|^2 + \left\| \frac{\partial \delta\mathbf{T}(X, Y; \Delta\mathbf{P})}{\partial Y} \right\|^2$$

The derivatives of $E_{data}(\Delta\mathbf{P})$ along $\Delta P^{[x]}_{(m,n)}$ can be calculated as

$$\frac{\partial E_{data}(\Delta\mathbf{P})}{\partial \Delta P^{[x]}_{(m,n)}} = -\iint\limits_{\Omega} 2r \frac{\partial r}{\partial \Delta P^{[x]}_{(m,n)}} dxdy \quad (9)$$

where

$$\frac{\partial r}{\partial \Delta P^{[x]}_{(m,n)}} = \iint_{(X',Y') \in bin(X,Y)} \varphi(\cdot) \frac{\partial \mathbf{T}(X',Y';\Delta\mathbf{P})}{\partial \Delta P^{[x]}_{(m,n)}} dX'dY' \quad (10)$$

The derivatives of $E_{smoothness}(\Delta\mathbf{P})$ along $\Delta P^{[x]}_{(m,n)}$ is calculated as

$$\begin{aligned}\frac{\partial E_{smoothness}(\Delta\mathbf{P})}{\partial \Delta P^{[x]}_{(m,n)}} &= 2\iint \frac{\partial \delta\mathbf{T}(X,Y;\Delta\mathbf{P})}{\partial X} \cdot \frac{\frac{\partial \delta\mathbf{T}(X,Y;\Delta\mathbf{P})}{\partial X}}{\partial \Delta\mathbf{P}^{[x]}_{(m,n)}} dXdY \\ &+2\iint \frac{\partial \delta\mathbf{T}(X,Y;\Delta\mathbf{P})}{\partial Y} \cdot \frac{\frac{\partial \delta\mathbf{T}(X,Y;\Delta\mathbf{P})}{\partial Y}}{\partial \Delta\mathbf{P}^{[x]}_{(m,n)}} dXdY \quad (11)\end{aligned}$$

where r in data term is defined as $\mathbf{I}_{LR}(x, y) - f(X, Y; \Delta\mathbf{P})$. The derivation for $\Delta P^{[y]}_{(m,n)}$ are similarly obtained.

Super-resolution methodology requires sub-pixel registration: SR reconstruction can be seen as "combining" new information from LR images to obtain a SR image. If the LR images have sub-pixel shifts from each other, it is possible to reconstruct a SR image. Otherwise, if the LR images are shifted by integer units, then each image contains the same information, and thus there is no new information that can be used to reconstruct a SR image. High frequency areas on a human face usually correspond to facial features such as eyes, eyebrows and mouth. During facial expression changes, these facial features deform the most among the face. In LR facial images, the fact is that these facial features have a few pixels, which are blurred from high resolution images and noisy. If we deform the LR image, the deformation can not capture the subtle movements. Moreover, interpolation in LR image during the process of deformation smooths out the high frequency features since they only occupy a few pixels. This leads to the loss of new information which can be used to reconstruct SR image.

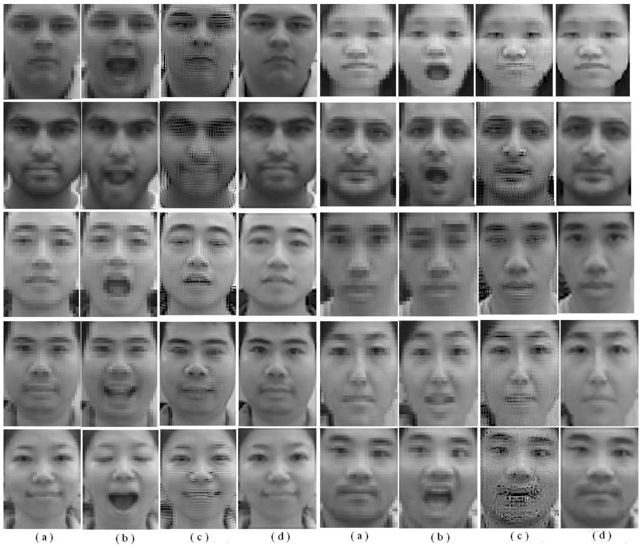

Fig. 3. Super-resolution Results using SR algorithms in [1].(a)-(b) Low-resolution images. (c)Reconstructed SR images using global registration. (d)Reconstucted SR images using global and our RAIFFD local deformation approach.

We use three SR algorithms [1][2][3].

3. EXPERIMENTAL RESULTS

Data and Parameters: We record 10 video sequences of 10 people with each lasting about 3 minutes. During the recording, each person was asked to look at the camera at a distance of about 20 feet from the camera and make continuous expression changes. The average size of the face is about 75 × 70 with maximum size at 115 × 82 and minimum size at 74 × 58. We blur these videos with a 5 × 5 Gaussian kernel and down-sample them into image sequences with average size 35 × 27. We then use our proposed approach to estimate the flow fields and super-resolve these LR images to acquire SR images. We found that 35 frames were enough to cover most of the expression changes the people made during recording. We then use 35 frames for each person for super-resolution.

Super-resolution Results on global registration vs. global + RAIFFD local deformation: We super-resolve the input LR images to acquire SR images using the SR algorithms [1]. The results for the complete data are shown in Fig. 3. The SR results in (d) are much better than the results in (c), especially in the high-frequency areas such as mouth, eyes and eyebrows. This is due to the reason that the global only alignment method can not capture the local deformations on these local parts when faces have expression changes. Our global+local approach not only finds the global transformation for the global motion, but also captures the local deformations.

In order to measure the performance of our algorithm, we compute peak signal-to-noise ratio (PSNR) as the measurement between original high-resolution and reconstructed SR images. Considering the fact that local motions mostly occur in the high-frequency areas such as eyes and mouth (see Fig. 4, we calculate the PSNR values only on the marked regions between the reconstructed SR image ((a) and (b)) and the original high-resolution image. The results are shown in Fig. 5. We find that PSNR of the areas for our global+local approach is much better than that for global only approach. The average PSNR value is 20.6261 for the global approach and 26.1327 for our global+local approach.

Similar to the above experiments using [1], we implement two methods [2][3] both on the globally aligned data and the data aligned using our global+local model. We find that the quality of SR images using [2] outperforms the method in [3]. This does not surprise us because the proposed method in [2] uses median estimator to replace the sum in the iterative minimization in [3]. This reduces the influence of the large projection errors due to inaccurate motion estimation.

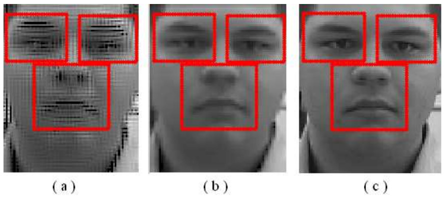

Fig. 4. Marked regions for calculating peak signal-to-noise ratio (PSNR). (a) Reconstructed SR image using global registration. (b) Reconstructed SR image using global + local deformation approach. (c) Original High-resolution image.

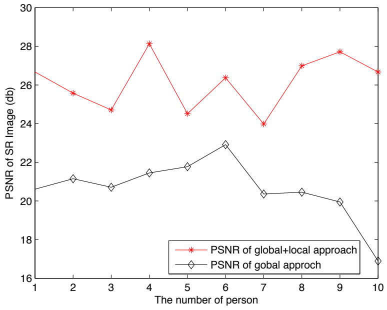

Fig. 5. Comparison of PSNR values between global reconstructed SR and global+local reconstructed SR images.

4. CONCLUSIONS

Reconstruction of SR facial images from multiple LR images suffers from the special characteristics of human face. Among these difficulties, non-rigidity of human face is a critical issue since accurate registration is an important step for SR. The experimental results show that the proposed SR framework can effectively handle the local deformations of human face and produce better SR image than the global approach.

5. REFERENCES

[1] 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.

[2] A. Zomet, A. Rav-Acha, and S. Peleg, "Robust super resolution," *Proc. of IEEE International Conf. on Computer Vision and Patern Recognition (CVPR)*, vol. 1, pp. 645–650, 2001.

[3] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy and undersampled measured images," *IEEE Trans. Image Processing*, vol. 6, pp. 1646–1658, December 1997.

[4] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 1167–1183, Sept. 2002.

[5] G. Dedeoglu et al., "High-zoom video hallucination by exploiting spatio-temporal regularities," *CVPR'04*, vol. 2, pp. 151–158, Jun. 2004.

[6] W. Liu et al., "Hallucinating faces: Tensor patch super-resolution and coupled residue compensation," *CVPR'05*, vol. 2, pp. 478–484, 2005.

[7] D.P. Capel and A. Zisserman, "Super-resolution from multiple views using learnt image models," *CVPR'01*, vol. 2, pp. 627–634, 2001.

[8] C. Liu et al., "A two-step aprroach to hallucinating faces: Global parametric model and local nonparametric model," *CVPR'01*, vol. 1, pp. 192–198, 2001.

[9] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025–1039, 1998.

[10] X. Huang, N. Paragios, and D. Metaxas, "Shape registration in implicit spaces using information theory and free form deformations," *IEEE Trans. on Medical Imaging*, vol. 28, no. 8, pp. 1303–1318, August 2006.