I.J. Image, Graphics and Signal Processing, **2013, 3, 1-9**
Published Online March 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2013.03.01

Facial Image Super Resolution Using Weighted Patch Pairs

Payman Moallem
Department of Electrical Engineering, University of Isfahan, Isfahan, Iran
e-mail: p_moallem@eng.ui.ac.ir

Sayed Mohammad Mostafavi Isfahani, Javad Haddadnia
Department of Electrical and Computer Engineering, Hakim Sabzevari University, Sabzevar, Iran
e-mail: mostafavi.isfahani@gmail.com, haddadnia@sttu.ac.ir

Abstract— A challenging field in image processing and computer graphics is to have higher frequency details by super resolving facial images. Unlike similar papers in this field, this paper introduces a practical face hallucinating approach with higher quality output images. The image reconstruction was based on a set of high and low resolution image pairs. Each image is divided into defined patches with overlapped regions. A patch from a defined location is removed from the low resolution (LR) input image and is compared with the LR patches of the training images with the same location. Each defined LR patch has a defined high resolution (HR) patch. Based on the Euclidean distance comparison, each patch of every single image in the training images database receives a specific weight. This weight is transferred to its relevant HR patch identically. The sum of the gained weights for one specific location of a patch is equal to unity. The HR output image is constructed by integrating the HR hallucinated patches.

Index Terms— Face Hallucination, Super Resolution, Image Patches, PSNR and SSIM

I. INTRODUCTION

The ever increasing demand on surveillance cameras is quite visible nowadays. Some major users of such technologies are the banks, stores and parking lots. When using security cameras or cameras recording far distances, we usually face low resolution - low quality facial images. These images do not have sufficient information for recognition tasks or other usages and require a boosted upgrade in their resolution quality. Because there are always the high frequency details that carry the typical information used in processing techniques. This upgrade is performed with different automatic or manual process.

Surveillance and monitoring systems like many other video based applications must extract and enhance small faces from a sequence of low-resolution frames [1], [2]. Direct interpolating of the input image is the simplest way to increase image resolution with algorithms such as cubic spline or nearest neighbor. But on the other hand the performance of direct interpolation is usually blurry and poor since no new information is added in the process.

The process of increasing the resolution in images is called Super resolution. If the image that is being super resolved is a human facial image, the process in this special domain is called face hallucination. The term "face hallucination" was first coined by Baker and Kanade [3-4]. This is a borrowed term which is originally a medical phrase that is used to describe a common symptom of severe mental disorder in patients who have visions or imaginary perceptions which does not exist. The purpose of this naming is that in Baker and Kanade's method, the output image will have high frequency facial hallucinated details that are not available in the LR image. Super resolution has diverse applications such as in medical images, aerial images, surveillance cameras expanding web images, improving old pictures, converting NTSC television images into HDTV and etc.

There are two main methods in resolving LR images. The first method is using reconstruction-based methods and the second one is using learning-based methods. The former method uses a sequence of images as the input and reconstructs an output image by the multiple inputs. The latter method uses learning, created by given previous given information from a database to form a higher resolution image from the input LR image.

The proposed method relies on the second method. Our main aim in face hallucinating is to reach a high quality super resolved image from the LR single frame input image. Since face detection is one of the important research issues by itself and needs its own algorithmic and conditional considerations it will be remained beyond the scope of this paper and the input images will all be previously extracted and cropped manually. Extensively surveyed papers on face recognition are presented by Yang et al. [5] and Kakumanua et al. [6]. The given cropped facial image is the input of the hallucination algorithm as shown in Fig. 1.

Copyright © 2013 MECS

Fig. 1 An example of a low resolution single frame facial image: a) cropped facial region b) cubic spline interpolating c) desired HR image.

Prior work on face hallucination will be discussed briefly in section 2. The creation of image patches and the overall proposed system is discussed in section 3. The steps on creating the learning database come after that and then there are experimental results. Finally the concluding texts are the ending sections.

II. Related Works

The first two to propose a "recognition-based" super-resolution (SR) method in their literature were Baker and Kanade [3-4]. In their pixel-wise method, it infers the high frequency components from a parent structure with the assistance of training samples. Their technique precisely functions by disintegrating an image into a pyramid of features including the first and second derivatives of the Gaussian pyramid and the Laplacian pyramid, and then searching the nearest neighbors of each pixel in this pyramid through a training dataset. In order to generate the target HR image, the high-frequency pyramid features for the input pixels are selected based on maximum a posteriori (MAP) formulation of the LR improvement problem.

A non-parametric patch-based prior along with the Markov random field (MRF) model to generate the desired HR images was executed by Freeman et al. [7]. By applying a homogeneous MRF model to learn the statistics between low-resolution "image" patches and underlying high-resolution "scene" patches, and the relation between neighboring "scene" patches, a framework for handling low-level vision tasks were implemented. The model is applied for generic image super-resolution when the scene and image are high- and low-frequency bands respectively.

Liu et al. [8] propose to integrate a holistic model and a local model for SR reconstruction. The performed two-step approach is executed by integrating a global parametric model with Gaussian assumption and linear inference and a nonparametric local model based on MRF. Both of the two methods used complicated probabilistic models and were based on an explicit resolution-reduction-function, which is sometimes unavailable in practice [9].

Inspired by a well-known manifold learning method, Locally Linear Embedding (LLE), Chang et al. [10] implemented the Neighbor Embedding algorithm based on the assumption that small patches in the low- and high-resolution images form manifolds with similar local geometry in two distinct spaces. The local distribution structure in sample space is preserved in the down-sampling process, where the structure is encoded by patch-reconstruction weights.

Wang and Tang [11] suggested a face hallucination method using principal component analysis (PCA) to represent the structural similarity of face images. They considered the face hallucination problem as a transformation between different face styles. The implemented method used PCA to fit an input low-resolution face image to a linear combination of low-resolution face images in the training set. A high-resolution image was then rendered by replacing the low-resolution training images with their high-resolution correspondences, while retaining the same combination coefficients. However, this method only utilizes global information without paying attention to local details.

Park et al. [12] proposed a technique derived from example-based hallucination methods and morphable face models. They proposed a recursive error back-projection method to compensate for residual errors, and a region-based reconstruction method to preserve characteristics of local facial regions. Then defined an extended morphable face model, in which an extended face is composed of the interpolated high-resolution face from a given low-resolution face, and its original high-resolution equivalent.

An example-based Bayesian method for 3D-assisted pose-independent facial texture SR was introduced by Mortazavian et al. [13] Bayesian framework was implemented to reconstruct a high-resolution texture map given a low-resolution face image. Their method utilizes a 3D morphable model to map facial texture from a 2D face image to a shape- and pose- normalized texture map and vice versa.

III. Weighted Patch Pairs Super Resolving

The process of creating HR for LR input images while adding high frequency components to the image is similar to solving a mathematical problem with "*m*" equations and "*n*" unknown components while the quantity of the unknown components compared to the quantity of available equations are much more. This process is almost impossible unless we moderate the problem by narrowing down the domain into a specific application which in this special case is the human face domain. Since face images are well structured and have similar appearance, they span a small subset in the high

Copyright © 2013 MECS

dimensional image space [14]. This similarity can be more obvious when comparing texture and color of the skin or the shape of the eyes, nose, lips and the space between them in human facial images within the same race, sex and age. Utilizing the faces structural resemblance, implies that the high frequency detail can be inferred from the low frequency components. On the other hand human facial features vary from one person to other especially depending on their age, sex and ethnicity. And even in the same mentioned conditions human faces have a great appearance variety. This makes using holistic and global methods to act very poor without having the ability to create certain specific high frequency components in the facial image which act as visible traits for human based and machine vision recognition. Hallucination face images by the aid of weighted image pairs patches helps to improve the resolution of such images without blurring the image or giving a mean global output.

A. *Creating the Patches*

The two dimensional input image will be converted into a vector of the pixels at the first stage. Each image pair in the database was created by one HR image with its low quality image that is the smoothed and down-sampled image of the HR pair. Both high and low resolution images are disintegrated into fixed size image patches with defined locations and with an overlapped region with their neighbor patches. Each LR image patch will be a pair for its corresponding HR patch. The ratio of the size of the LR patch to HR patch is proportional to the ratio of the HR image to the LR image. Each patch will be reshaped into a vector for further process. Each facial image will be considered as a matrix like $\{X^M(i,j)\}_{M=1}^N$ made of overlapping image patches where N is the total number of patches in image X. The patch in row i and column j in the patch matrix is mentioned by $X^M(i,j)$. Assume that each square patch fills $n\times n$ pixels. For the LR image $\{X^M(i,j)\}_{M=1}^N$ if n is an odd digit, the patch $X_L^M(i,j)$ will be overlapping $n\times[(n\text{-}1)/2]$ pixels with its neighbor patch and for its proportional HR image the patch $X_H^M(i,j)$ which fills $qn\times qn$ pixels will have overlapping on $(qn)\times[q(n\text{-}1)/2]$ pixels. And if n is an even digit, the patch $X_L^M(i,j)$ will be overlapping $n\times(n/2)$ pixels with its neighbor patch and for its proportional HR image the patch $X_H^M(i,j)$ which fills $qn\times qn$ pixels will have overlapping on $qn\times(qn/2)$ pixels. An example of the creation of facial patch is shown in Fig. 2.

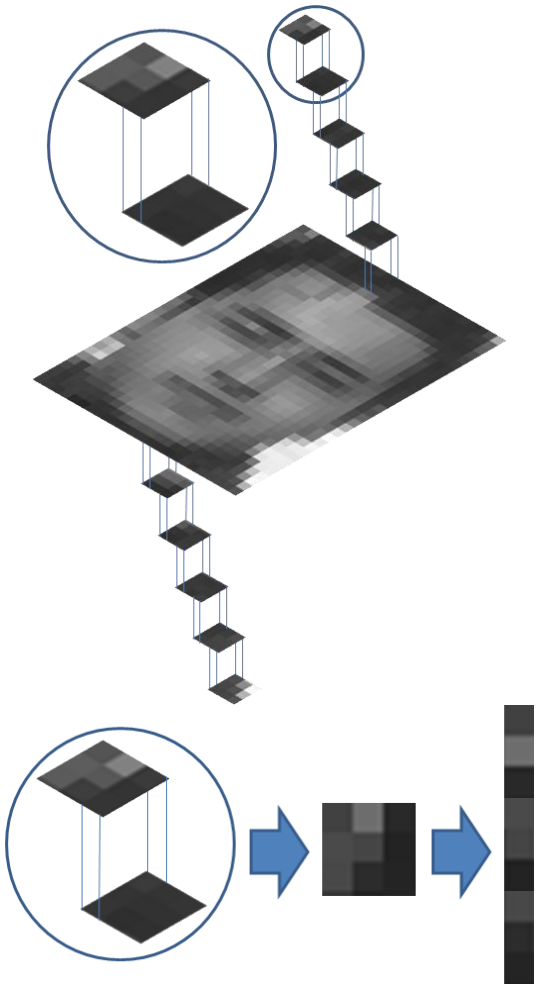

Fig. 2 Creation of the patches from the facial image

B. *Reconstruction Based on Patch*

Assume that X^p stands for the available learning database images having p=1,2,…,P where P is the number of these images. Each patch in the learning database will be in the form of $\{X^{pM}(i,j)\}_{M=1}^N$.

The reconstructions weight matrix will be as $w_p(i,j)$ where $w_p(i,j)$ is the representative for each patches share in the (i,j) location for reconstructing the input image located in the same mentioned location.

Each mentioned weight in its (i,j) location was considered in a way that the sum of the total weights will become unit. Suppose that Y is the image which will be super resolved and consider Y as patches in the form of $\{Y^M(i,j)\}_{M=1}^N$. For the images in the learning database, the images available at the location (i,j) like $X^{pM}(i,j)$, will be a placed patch for the image patch of $Y^M(i,j)$. We expect that each patch like $Y^M(i,j)$ in the human face image of $\{Y^M(i,j)\}_{M=1}^N$ to be expressed as (1).

$$Y^M(i,j)=\sum_{p=1}^{P} w_m(i,j)X^{pM}(i,j)+e \tag{1}$$

Copyright © 2013 MECS

where *e* remarks the reconstruction error. From (1) we can realize that the optimum weights for reconstructing the image depends on minimizing the quantity of the error *e*.

$$w(i,j)=\arg\min_{w_p(i,j)}\left\|Y^M(i,j)-\sum_{p=1}^{P}w_m(i,j)X^{pM}(i,j)\right\|^2 \quad (2)$$

where $w(i,j)$ is a P dimensional weight vector for each reconstruction weight of $w_p(i,j)$. For each value of p=1,2,…,P the value of U can be calculated by (3).

$$U=Y^M(i,j).A^T-X \quad (3)$$

where A is a column vector of ones and U is matrix which its columns are the values of the $X^{pM}(i,j)$ patches. The local matrix of V can be found by $V=U^TU$. Equation (2) is a constrained least squares problem where its solution can be as:

$$w(i,j)=\left(V^{-1}A\right)/\left(A^TV^{-1}A\right) \quad (4)$$

A better way in obtaining $w(i,j)$ is solving the linear system $V.w(i,j) = A$ followed by changing the scales in order to a sum of one in the patches weights. $w(i,j)$ can be used to reconstruct the new image patch $Y_R^M(i,j)$ from (5) as:

$$Y_R^M(i,j)=\sum_{p=1}^{P}w_p(i,j)X^{pM}(i,j)\cong Y^M(i,j) \quad (5)$$

where $Y_R^M(i,j)$ is a vector converted into a matrix used to construct the whole image. The whole image is constructed by integrating the patches considering their original locations. The values of the overlapping pixels are calculated by averaging the pixel values between the two overlapping neighbor patches.

C. Super Resolving Based on Patches

While having the LR image Y_L and the HR pair of this image Y_H which is q^2 times bigger, the image resampling can be expressed by [3]:

$$Y_L=\frac{1}{q^2}\sum_{k=0}^{q-1}\sum_{l=0}^{q-1}Y_H(qi+k,qj+l)+n(i,j) \quad (6)$$

where q is a positive number and $n(i,j)$ represents a random noise. If Y_L, Y_H and n are vectors as $L\times1$, $K\times1$ and $K\times1$ respectively, (6) can be simplified to:

$$Y_L=HY_H+n \quad (7)$$

where H is a $K\times L$ matrix. Equation (7) includes a smoothing and down-sampling process which can be expressed as (8) in image patches.

$$\left\{Y_L^M(i,j)\right\}_{M=1}^{N}=H\left\{Y_L^M(i,j)\right\}_{M=1}^{N}+n \quad (8)$$

For each $Y_L^M(i,j)$ patch in the LR input image the weight of $w_p(i,j)$ is obtained from (1) to (4) and satisfies (9):

$$Y_L^M(i,j)\cong\sum_{p=1}^{P}X_L^{pM}(i,j)w_p(i,j) \quad (9)$$

Replacing each LR image patch $X_L^{pM}(i,j)$ with its proportional HR one $X_H^{pM}(i,j)$ results in (10).

$$\sum_{p=1}^{P}X_L^{pM}(i,j)w_p(i,j)=Y_L^{pM}(i,j) \quad (10)$$

Without considering the effect of noise from (7) and (10) we have:

$$\begin{aligned}H.Y_H^{pM}(i,j)&=\sum_{p=1}^{P}H.X_H^{pM}(i,j)w_p(i,j)\\&=\sum_{p=1}^{P}X_L^{pM}(i,j)w_p(i,j)\end{aligned} \quad (11)$$

From (9) and (11) we have:

$$Y_L^M(i,j)\cong H.Y_H^M(i,j) \quad (12)$$

It can be seen from (12) that the degradation value of $Y_H^M(i,j)$ is close to the LR input $Y_L^M(i,j)$. All the HR patches $Y_H^M(i,j)$, are integrated to reconstruct the HR whole final output image $\left\{Y_H^M(i,j)\right\}_{M=1}^{N}$ considering their original locations. The values of the HR pixels which are in the overlapping regions are calculated by averaging the pixel values between the two overlapping HR neighbor patches.

The symbolic super resolving process shown in Fig. 3 can be described in these steps. At the first step the input image will be reshaped into patches as described before. In the next step each LR patch is compared to the proportional LR patches in the learning database according to their location and receives one weight by each comparison. In the third step the concluded weights created by comparing the LR images are used as weights proportionally on the HR pairs and at the end the HR output is created by integrating the HR output images created in the third step.

Copyright © 2013 MECS

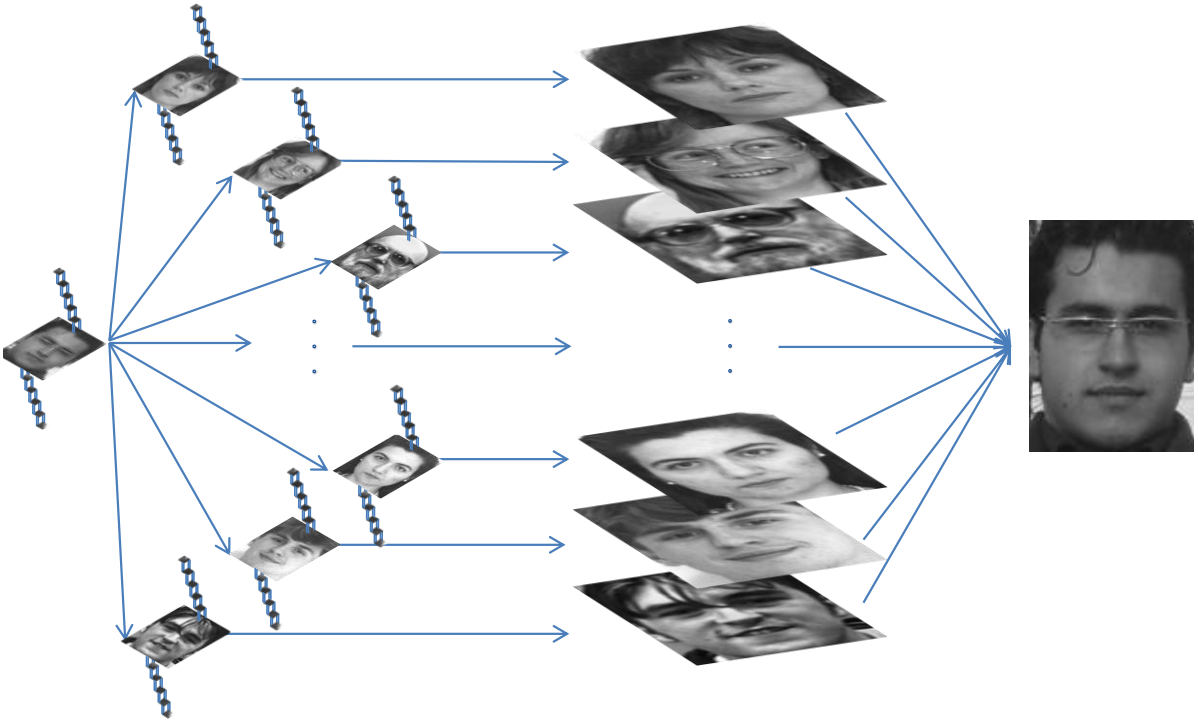

Fig. 3 The framework of using high and low resolution image pairs in the proposed face hallucination method

D. The Learning Database

In order to create the learning image database 600 frontal face images were used. These images were brought from famous available image databases such as ORL, ESSEX, YALE, ABERDEEN and LFW. Using multiple databases increases the functionality and makes the system robust. Although different available races, illuminations, facial poses and other conditions that has effects on the image reconstruction are gathered together in the face database, the proposed system functions hale.

The database images were all firstly converted to gray-scale. In order to create the high-LR image pairs from the available databases at first the images which the region containing the face had a resolution higher than 96 × 128 pixels were selected. After the face region was centered and replaced in a 96 × 128 window in such a way that the center of the two eyes and the ending point of the noise followed a pre marked face image used as a reference it was cropped. Following this operation the LR image pair was created by having a 4x size reduction which was handled by smoothing and down-sampling the HR pair ending in a 24×32 pixels facial image. The mentioned window sizes are determined to 96×128 and 24×32 pixels because most automated face processing tasks are possible with 96×128 pixel images [3].

We developed a resolution enhancement in a 24×32 pixels facial image. The mentioned window sizes are determined to 96×128 and 24×32 pixels because most automated face processing tasks are possible with 96×128 pixel images [3]. We developed a resolution enhancement algorithm, specifically for faces, that can convert a small number 24×32 pixel image of a face into a single 96×128 pixel image. It shall be mentioned using this method we can still have an increase in resolution from a desired input image size to a desired output by changing the database image pairs but based on the image pair sizes the quality of super resolving may vary to a higher or lower facial image output.

IV. Evaluation

In order to quantifiably measure the performance of the proposed algorithm, we evaluated the peak signal-to-noise ratio (PSNR) and the Structural SIMilarity (SSIM) index [15] as measurements between the ground truths face images and the hallucinated facial images. The Mean Square Error (MSE) and the PSNR are the two error metrics used to compare image compression quality which have clear physical meanings and are simple to compute. The MSE represents the cumulative squared error between the compressed and the original image, whereas PSNR represents a measure of the peak error. The lower the value of MSE means the lower the error. Again, the higher the PSNR also means the better the quality of the compressed or reconstructed image. However, they are not very well matched to perceived visual quality. To compute the PSNR, at first the mean-squared error is calculated using (13):

$$MSE = \frac{\sum_{M,N}[I_1(m,n) - I_2(m,n)^2]^2}{M * N} \quad (13)$$

where M and N are the number of rows and columns in the input images, respectively. Followed by (13) PSNR is computed using (14):

Copyright © 2013 MECS

$$PSNR = 10\log_{10}\left(\frac{L^2}{MSE}\right) \quad (14)$$

where L is the maximum fluctuation in the input image data type. Since the facial images are gray-scale 8-bit unsigned integer data type, L is 255.

The MSE/PSNR is not as consistent as the human eyes perception in measuring the visual quality of images; hence the SSIM as a complementary measurement to evaluate the respective methods performances is also adopted. 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 separates 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 perceptual image quality. For SSIM, the highest score happens when two images are identical. The PSNR and the SSIM values for a sample of randomly selected individuals are plotted in Fig. 6 and Fig. 7 respectively. Fig. 6 and 7 shows that our proposed method has better score than that of the other two techniques in terms of PSNR and SSIM, which were confirmed also by the visualization of Fig. 4

(e) (d) (c) (b) (a) (e) (d) (c) (b) (a)

Fig. 4. Comparing the proposed method a)low resolution facial image b)cubic spline c)Baker's method d)proposed method e)original HR image

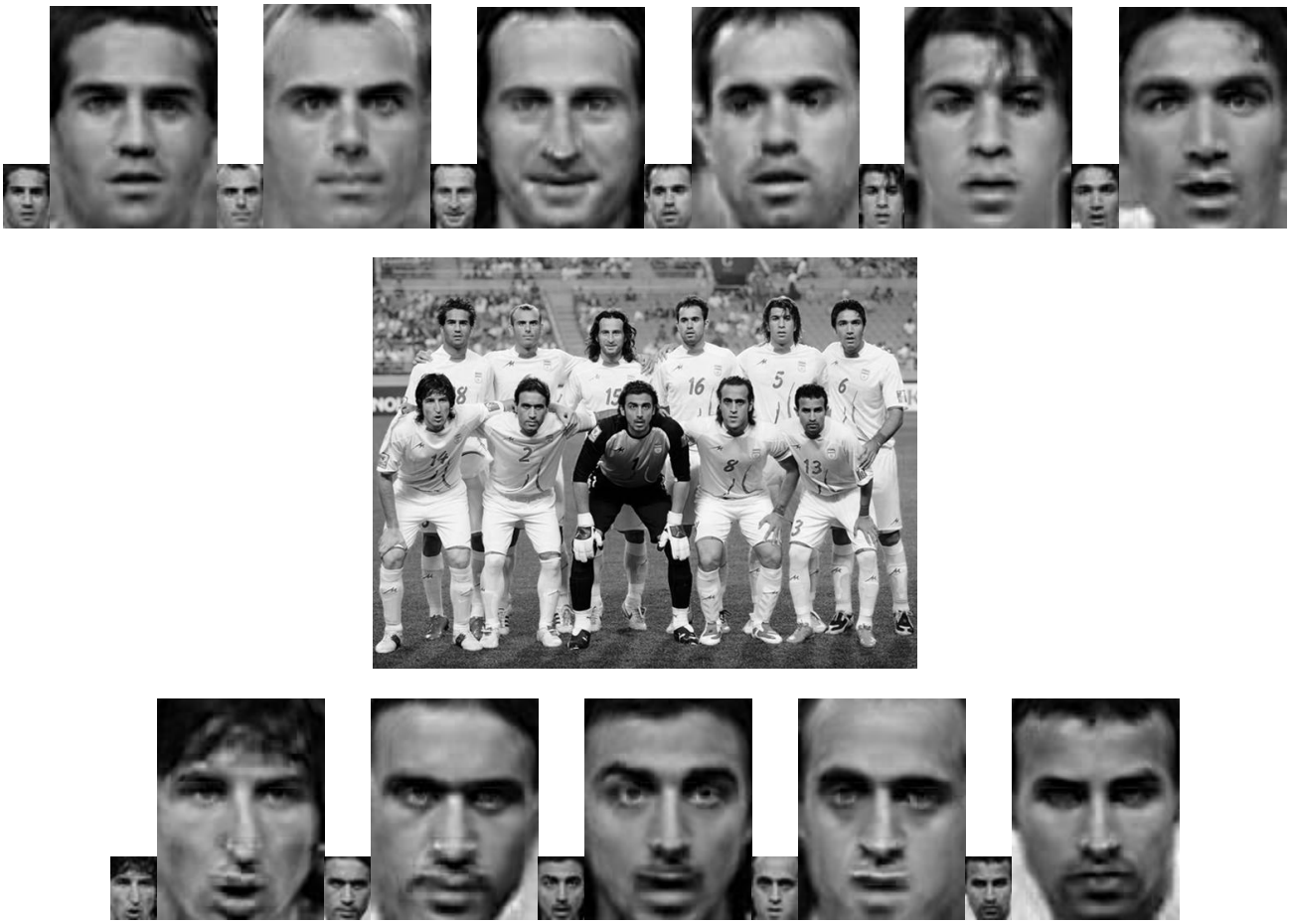

Fig. 5 A real world image of Irans soccer team 2010 with the hallucinated facial images based on the proposed method

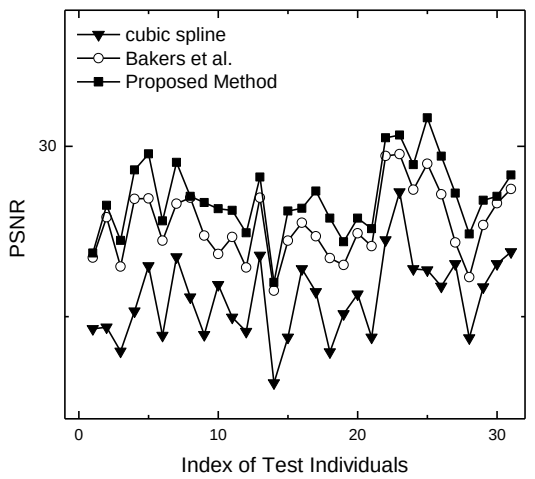

Fig. 6 PSNR values of the hallucinated results of three different methods of 31 randomly selected individuals

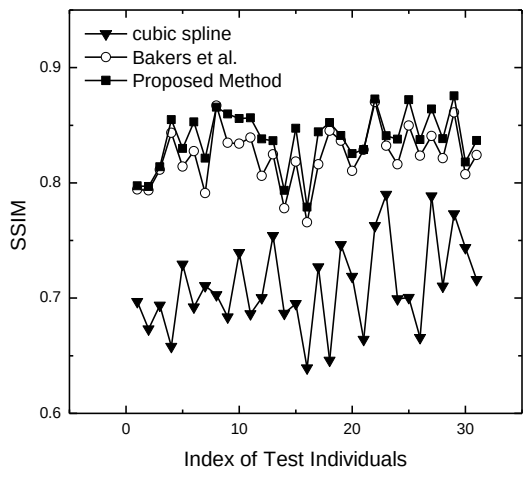

Fig. 7 SSIM values of the hallucinated results of three different methods of 31 randomly selected individuals

Copyright © 2013 MECS

V. Experimental Results

In order to show the improvement of the images resolution and having an easier comparison, the results of implying the described method is brought in Fig. 4. At first the LR input image with the size of 24 × 32 pixels is shown the result of face hallucinating using cubic spline and Bakers [3] method is brought next and finally the mentioned method and the original HR image is brought in the last two columns. A real world image is brought in Fig. 5 including the LR images used as inputs and the hallucinated HR outputs. In order to quantifiably measure the performance of the proposed algorithm, we evaluated the peak signal-to-noise ratio (PSNR) and the Structural SIMilarity (SSIM) index as measurements between the ground truth face images and the hallucinated facial images brought in Fig. 6 and 7. Comparing Fig. 4, 6 and 7, our proposed method has better score than the other two techniques in terms of PSNR and SSIM and also by the means of human visualization.

VI. Conclusion

A practical facial super resolution method based on the acquired weights form image patch pairs was described in this paper. High and low resolution image pairs in the image database were the key for in the learning stage. A patch was removed from a defined location of the LR input image and was compared with the LR training images patches with the same location. By this comparison a weight for each LR images patch is gained. After estimating the corresponding weights, the learning databases HR images patch which was a pair for the LR images patch was used to hallucinate the input LR image. The HR output image was constructed by integrating the HR hallucinated patches. Experimental results demonstrate performance superiority and novelty in terms single modal face hallucination.

References

[1] B. Tom and A. Katsaggelos, "Resolution enhancement of monochrome and color video using motion compensation," IEEE Trans. Image Process., vol. 10, no. 2, pp. 278–287, Feb. 2001.

[2] W. Liu, D. Lin, and X. Tang, "Neighbor combination and transformation for hallucinating faces," in Proc. IEEE Conf. Multimedia and Expo, Amsterdam, The Netherlands, Jul. 2005, pp. 145–148.

[3] S. Baker and T. Kanade, "Hallucinating Faces," in Proc. of Inter. Conf. on Automatic Face and Gesture Recognition, pp. 83-88, 2000.

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

[5] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 1, pp. 34–58, Jan. 2002.

[6] P. Kakumanua, S. Makrogiannisa, and N. Bourbakis, "A survey of skin-color modeling and detection methods," Pattern Recognit., vol. 40, no. 3, pp. 1106–1122, Mar. 2007.

[7] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-Based Super-Resolution," IEEE Comput. Graph. Appl., vol. 22, no. 2, pp. 56-65, 2002.

[8] C. Liu, H. Shum, and C. Zhang, "A Two-Step Approach to Hallucinating Faces: Global Parametric Model and Local Nonparametric Model," in Proc. of CVPR, Vol. 1, pp. 192- 198, 2001.

[9] W.Liu,D.H.Lin,X.O.Tang,Hallucinating faces: Tensor Patch super-resolution and coupled residue compensation,in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,United States, 2005,pp.478–484.

[10] H. Chang, D. Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in Proc. IEEE Conf. Computer Vision and *Pattern Recognition*, 2004, pp. 275–282.

[11] X. Wang and X. Tang, "Hallucinating face by eigentransformation," to appear in IEEE Trans. on Systems, Man, and Cybernetics, Part-C, Special issue on Biometrics Systems, 2005.

[12] J. S. Park and S. W. Lee. An example-based face hallucination method for single-frame, low-resolution facial images. IEEE Transactions on Image Processing, 17(10):1806–1816,2008.

[13] P. Mortazavian, J.V Kittler, W.J. Christmas, "3D-Assisted Facial Texture Super-resolution", BMVC09 xx-yy 2009

[14] P. S. Penev, and L. Sirovich, "The Global Dimensionality of Face Space," *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 264-270, 2000.

[15] 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.

Copyright © 2013 MECS

Payman Moallem, male, is an Associate Professor at University of Isfahan, Iran. His research interests include computer vision, digital image and video processing and artificial intelligence.

Sayed Mohammad Mostafavi Isfahani, male, received his MSc in Electrical Engineering, from Hakim Sabzevari University, Iran. His research interests include digital image processing, machine vision, circuits & systems, artificial intelligence and robotics.

Javad Haddadnia, male, is an Associate Professor at Hakim Sabzevari University, Iran. His research interests include biomedical image processing, neural networks, Fuzzy system and data mining.

Copyright © 2013 MECS