LEARNING LOCAL PIXEL STRUCTURE FOR FACE HALLUCINATION

Yu Hu\textsuperscript{1, 2}, Kin Man Lam\textsuperscript{2}, Guoping Qiu\textsuperscript{3}, Tingzhi Shen\textsuperscript{1}, Hui Tian\textsuperscript{1}

1. Department of Electronic Engineering, School of Information Science and Technology, Beijing Institute of Technology, Beijing 100081, P. R. China
2. Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong.
3. School of Computer Science, University of Nottingham, UK

ABSTRACT

In this paper, we present a novel learning-based face hallucination method based on the assumption that similar faces will have similar local pixel structures. We use the low-resolution (LR) input face to search a database for K example faces that are the most similar to the input and align them with the input accordingly. The local pixel structures of the target high-resolution (HR) image are learned from those warped HR example faces in a neighbor embedding manner, and a total variation (TV) constraint is employed to aid the learning of all pixels’ embedding weights. The learned local pixel structures are then used as constraints to reconstruct a HR version of the input face. Experimental results show that the method performs well in terms of both reconstruction error and visual quality.

Index Terms—face hallucination, local pixel structure, TV norm, super resolution

1. INTRODUCTION

Learning-based facial image super-resolution (SR) is also known as “face hallucination” which refers to the use of learning or recognition based methods to reconstruct a HR facial image from its LR observation. This kind of method owns the advantages that they can generate HR facial details more effectively and can achieve a higher magnification factor than traditional reconstruction-based SR method. Baker et al. [1] proposed the “recognition” based SR method using a MAP formulation. The method constructs a group of multi-scale feature pyramids using training images, and searches the high-frequency components for the LR input through these feature pyramids. In [2], Qiu predicts a HR image from its LR counterpart based on an inter-resolution look-up table. This table captures the intrinsic correlation between the HR and LR image pairs, and is designed using vector quantization (VQ). The example-based SR approach by Freeman et al. [3] is another typical work in the learning-based SR field. The method employs a non-parametric patch-based prior along with the Markov random field (MRF) model to generate the target HR image. After their work, a number of patch-based SR methods based on the spirit of [3] have been shown to deliver improved performances [4].

All the abovementioned SR methods can be applied to both facial and natural images, depending on the types of images used in the training dataset. In contrast, the PCA-based SR algorithms are more specific for the facial image hallucination task. In [5], Wang et al proposed an eigentransformation method to super resolve a single LR facial image. They use the PCA technique to represent a LR input as a linear combination of the LR training samples. The desired HR image is then obtained by keeping the combination coefficients invariant and replacing the LR training samples with the corresponding HR counterparts. As the global facial structures are utilized in the eigentransformation method, it is very efficient to infer the HR facial components for the LR image. Some other promising works have also been done by researchers based on the similar technique, such as [6].

In this paper, we present a novel face hallucination method which uses local pixel structures to constrain the HR face image reconstruction. Since faces are regular patterns, it is reasonable to assume that similar faces should have similar local pixel structures. Based on this rationale, our new method first uses the input LR face to search a database for faces that are similar to the input, and then uses the HR images of those similar faces to learn the local structures of the HR image of the input face. The method captures the local pixel structure in the form of neighborhood embedding where a pixel is represented as a linear combination of its neighbors. A total variation (TV) norm constraint has been employed to aid the learning of neighbor embedding weights. Since the new method learns the local pixel structures from similar HR images and reconstructs the target HR face in the pixel level, it can generate good facial details in the estimated HR face image. We will compare the performance of our method with those described in [3, 5] in terms of both reconstruction error and visual quality, and show that our new method consistently gives better performances.

978-1-4244-7993-1/10/$26.00 ©2010 IEEE 2797 ICIP 2010
2. OVERVIEW OF THE PROPOSED SR TECHNIQUE

The standard super resolution (SR) problem for a single LR image normally can be described as the following image degradation process:

\[ I_I = D B I_h + n, \]

where \( I_I \) and \( I_h \) are the column vectors representing LR observation and target HR image respectively. \( B \) is the blurring matrix, and \( D \) is the decimation matrix. Here, we use the Gaussian type point spread function (PSF) for the generation of blurring matrix \( B \), and assume that the additive noise \( n \) is of zero mean white Gaussian noise. Thus, the Maximum Likelihood (ML) estimation for the target HR image \( I_h \) in equation (1) can be realized by minimizing the following cost function:

\[ \hat{I}_h = \arg \min_{I_h} \left\| DB\hat{I}_h - I_I \right\|^2, \]

where \( \hat{I}_h \) is the estimated HR image. Because this is an ill-posed inverse problem, there are many possible solutions for \( I_h \) and most of them will be spurious. To achieve a reasonable estimation for the target HR image, we incorporate a local pixel structure constraint into (2) to generate the desired HR pixels’ values in the target image. The constraint can be mathematically expressed as follows:

\[ \hat{I}_h(x) = \sum_{p \in C} \hat{I}_h(x + p) \cdot w_p(x), \]

where \( p \) is the displacement with respect to the pixel location \( x \) within a predefined neighborhood \( C \) (in this paper we set \( C \) as a 3x3 neighborhood window), and \( w_p(x) \) is the embedding weights which reflect the local pixel structures of \( \hat{I}_h \). Thus, eq. (3) stipulates that each target HR pixel in the estimated HR image and its neighbors should satisfy a specific linear combination relationship. This linear combination is entirely controlled by the embedding weights \( w_p(x) \), and can exhibit various properties in reconstructing the HR local primitives, such as the determination of directions or sharpness of the reconstructed edges. By assigning appropriate values for the embedding weights \( w_p(x) \) of each target HR pixel, we can reconstruct the target HR image by using this pixel structure constraint as a prior and incorporating it with the LR observation constraint in (2). The reconstruction procedure can be described as the following minimization procedure:

\[ \hat{I}_h = \arg \min_{I_h} \left\| DB\hat{I}_h - I_I \right\|^2 + \gamma \sum_{x} \left( \sum_{p \in C} \hat{I}_h(x + p) \cdot w_p(x) \right)^2, \]

The problem described by (4) is essentially a MAP estimation process. The two error terms, i.e. the LR observation constraint and the local pixel structure constraint, can be viewed as the data term and the prior term, respectively. Note that it is normally difficult to get an \( \hat{I}_h \) that can perfectly satisfy the equality in (3) when it is subjected to the LR observation constraint. So, we turn to find a solution that can best fit the equality in (3) by minimizing the total squared errors between the two sides of (3) for all pixels in the target image, as shown in (4). The contributions of the two terms, i.e. the data term and the prior term, are balanced by the regularization parameter \( \gamma \). To solve the above minimization problem, the embedding weights \( w_p(x) \) should first be determined. In this paper, we employ a training dataset of HR-LR facial image pairs to learn the embedding weights for the target HR pixels. The overall learning and reconstruction process is shown in Fig. 1.

For a given LR face image, we first search the training dataset for those faces that are most similar to the input image. Each training face in the database is stored in pairs, one LR and the corresponding HR. The LR version is used to compute the similarity with the input. The HR images of the \( K \) most similar faces are then used to estimate the embedding weights for the target HR face. The similarity between the face images is computed by measuring the Euclidean distance in the PCA subspace which is constructed by the first 200 principle components of the training dataset in this paper. The rationale for this scheme is that if two faces are similar, then they should have similar local structures as represented in the embedding weights of (3), which in turn means that the local structures of an unknown HR face image can be estimated from similar example faces. Because all faces have similar patterns, this is a reasonable assumption.

To make the idea work in practice, we need to warp each of the selected HR face examples to the LR input to get a better match. The warping method used in this paper is optical flow [7], and other warping methods such as the active appearance model can also be used. To generate a flow field of the HR grid, we first expand the LR versions of the example faces as well as the LR input into the desired HR size, and then take them as the input of the optical flow algorithm. The obtained HR flow field is used to warp the example HR face image accordingly. When \( K \) warped HR example faces have been obtained, we can then learn the desired local pixel structures of the target HR image from them (details will be given in Section 3). The learned local structures or embedding weights are used in (4) to derive the final reconstruction result.

Fig. 1. The diagram of the proposed face-hallucination method.
3. TOTAL VARIATIONAL LOCAL STRUCTURE LEARNING

Once all the example faces have been warped and aligned with the input face, we can learn the embedding weights. It turns out this is a crucial step to the success of deriving a good SR result. We employ the total variational (TV) method to aid the learning of desired local pixel structures. The choice of TV norm has found favor in the image restoration community because of its edge preserving property. As opposed to the $L_2$ norm, it does not over-smooth salient image primitives. For the SR problem here, we exploit its property in the estimation of local pixel structures of the target HR image, so as to suppress the noise occurred in the estimation and to preserve the desired edge structures. Here, the noise represents the irregularities appeared in the estimated local pixel structures, which are caused by the mismatch among respective HR example faces. As the perfect alignment is unavailable, this noise is inevitable.

We form a $K$-dimensional column vector $h_x = \left[ I_1^k(x), \ldots, I_K^k(x) \right]$ for each pixel location $x$, where $I_k^k(x)$ represents the pixel value of the $k^{th}$ warped HR face example at location $x$. We establish the relations between $h_x$ and its neighbours $h_{x,p}$ in a neighbour embedding manner, and find $w_p(x)$ by minimizing the following objective function:

$$\arg\min_{x} \sum_k h_k - \sum_p h_{x,p} \times w_p(x) + \frac{\lambda}{2} \sum_p \sum_x \nabla w_p(x) \cdot \nabla w_p(x)$$

where $N$ is the set of index for all pixel positions in the HR image. $\lambda$ is the parameter which balances the contributions of the two terms in (5). The first term measures the reconstruction errors between $h_x$ and the weighted combination of its neighbours $h_{x,p}$. This reflects the amount of difference with representing pixel $x$ as the specified linear combination of its neighbours. And the second term is a well-known TV norm constraint, which enforces a smoothness property on the variations of estimated embedding weights $w_p(x)$. By minimizing the objective function in (5), we can obtain the desired embedding weights $w_p(x)$ which reflects the common local structures of $K$ warped HR face examples at each location $x$. These embedding weights are directly used to approximate those of the target HR image, and are then used in (4) to generate an estimation of the HR image for the LR input.

It is worthy mentioning that the introduction of the TV norm constraint in (5) is important to the estimation of $w_p(x)$. When an insufficient number of HR face examples is selected for the LR input (for example, $K$ is smaller than the size of $C_1$, or when strong co-linearity exists among $h_x$ and its neighbours $h_{x,p}$, the problem of learning the embedding weights $w_p(x)$ will become ill-posed if we only employ the first constraint term in (5). By minimizing the TV norm of the embedding weights to be estimated, we can impose an extra smoothness constraint on the variations of the embedding weights, thus rendering the learning problem as well-posed. To solve the minimization problem in (5), a time-dependent partial differential equation (PDE) is usually established [8]:

$$\frac{\partial w_p(x)}{\partial t} = \text{div} \left( \frac{\nabla w_p(x)}{\sqrt{\nabla w_p(x)}} \right) - \lambda (h_x - \sum_p h_{x,p} \times w_p(x))^T h_{x,p}.$$  

When the PDE reaches its steady state, the solution is obtained. We adopt a level-set method as described in [8] to solve (6). The final reconstruction result for the target HR image is obtained by minimizing (4), and it is solved by using a standard gradient-based method.

4. EXPERIMENTAL RESULT

A dataset of 230 facial images of different individuals is used in our experiment. These images are extracted from the GT [9], AR [10], and FERET [11] databases, and are of mainly normal expressions and all frontal views. All the images are roughly aligned according to the two eye’s positions, and are without any accessories. The intensities and contrasts of images are also normalized. The size of the original HR image is 124×108, and the LR images are blurred and down-sampled from the HR counterparts to a size of 31×27, i.e. the magnification factor is 4. The standard deviation of Gaussian blurring kernel is 1.2.

We use a “leave-one-out” methodology to evaluate the performance of our SR method. At each round, one image is selected for testing and the remaining is used as the training dataset. We compare our method with the typical works in the patch-based and PCA-based SR methods as proposed in [3, 5] to show the effectiveness of our pixel-based local structure constraint in reconstructing the HR facial images. The number of HR example faces $K$ is set as 6. The values of $\lambda$ and $\gamma$ are found empirically to yield a satisfactory result. We tested respective algorithms for 100 facial images, and some example results are shown in Fig. 2. The reconstructed nose regions of the first and second individuals are also magnified for better comparison.

The results show that our method can have a much better visual performance than Freeman’s method can. This may be attributed to the fact that our method reconstructs each HR pixel of the target image by using the specifically learned local structures from HR example faces, and thus can better recover the detailed information in the image. In addition, the use of the TV norm constraint can suppress unwanted artifacts in the reconstruction process, which helps in producing clearer facial details. Wang’s method can generate plausible HR facial images. However, its result is somewhat different from the original image, and thus a lower fidelity level result. The quantitative measurements for the respective methods are tabulated in Table 1, in terms of both PSNR and SSIM [12]. Fig. 3 shows the MSEs of all
reconstructed facial images using the respective methods. Again, we can see that our method consistently outperforms the other two SR methods.

5. CONCLUSIONS

In this paper, we have developed a face-hallucination method which super-resolves a LR facial image using the local pixel structure as a constraint. The pixel structures are represented as the embedding weights of pixels with respect to their neighbors, and are learned from example HR faces whose LR counterparts are similar to the input LR image. An effective TV norm constraint has been employed to aid the learning of local pixel structures. Since the high-frequency information is learned from the reference face examples and captured in the local pixel structures, our method is capable of generating plausible facial details on the pixel level. Experimental results show that our algorithm can produce good hallucination results in terms of both reconstruction errors and visual quality. Future work on our method will include its extension to more generalized natural images with a patch-based framework.

6. REFERENCES

[12] Z. Wang, and A. C. Bovik, “Mean squared error: Love it or leave it?

TABLE I. THE AVERAGE PSNR AND SSIM OF THE RECONSTRUCTED HR IMAGES USING DIFFERENT ALGORITHMS.

<table>
<thead>
<tr>
<th>Average quality measure</th>
<th>PSNR (dB)</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic interpolation</td>
<td>22.54</td>
<td>0.71</td>
</tr>
<tr>
<td>Wang’s method</td>
<td>25.68</td>
<td>0.72</td>
</tr>
<tr>
<td>Freeman’s method</td>
<td>26.31</td>
<td>0.79</td>
</tr>
<tr>
<td>Proposed method</td>
<td>27.23</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Fig. 3. The MSEs of SR facial images of the respective methods.

Fig. 2. The HR images derived from: (a) the original HR images, (b) bi-cubic interpolation, (c) Wang’s method, (d) Freeman’s method, and (e) our proposed method. (f) The magnified nose regions of the 1st and 2nd individuals in (a) – (e) (from left to right), respectively.