Abstract
As the face image captured by a surveillance camera is typically very low-resolution (LR), blurred and noisy, traditional neighbor embedding method considers only one manifold (the LR image manifold) and fails very often to reliably estimate the intention geometrical structure. In this paper, we introduce the notion of neighbor embedding from the LR image manifold and the high-resolution (HR) one simultaneously and propose a novel neighbor embedding model, termed the coupled-layer neighbor embedding (CLNE), for surveillance face hallucination. CLNE differs substantially from other neighbor embedding models in that the former has two layers: the LR layer and the the HR layer. The LR layer in this model is the local geometrical structure of the LR patch manifold, which is characterized by the reconstruction weights; the HR layer in this model is a set of HR training patches that guide the K-nearest neighbor (K-NN) searching and geometrically constrain the reconstruction weights. By this coupled constraint paradigm between the adaptation of the LR layer and the HR one, CLNE can achieve a more robust neighbor embedding through the significant degradation process. Indeed, the experimental results confirm that our method outperforms the related state-of-the-art methods by having better objective values as well as better visual results.

Index Terms—Intelligence surveillance video, image processing, super-resolution, face hallucination, manifold learning

1. INTRODUCTION
Face hallucination (or super-resolution) refers to the process by which a higher-resolution enhanced face is reconstructed from one or more low-resolution (LR) images [1–8]. It benefits a number of real-world applications from image compression to face identification. Especially in intelligence surveillance videos, a higher-resolution face image with detailed features will be absolutely significant to raise the system’s performance [9–11]. Practical face hallucination methods may make use of a single still image or a sequence of consecutive video frames with sub-pixel translation for synthesizing a higher-resolution image [12]. In this paper, we focus on the problem of generating a high-resolution (HR) image from a single LR frontal image of a human face and inferring some high-resolution details missing from the original image that cannot be achieved by simply sharpening.

In recent years, there has been a good deal of research into learning-based approaches for face hallucination. These learning-based methods share the common characteristic of using a training set of HR and LR image pairs to build a co-occurrence model [13]. With the learnt model, one can then predict the missing details of the observed input LR image through borrowing information from some similar examples (patches) in the training set.

Inspired by locally linear embedding (LLE) [14] manifold learning, many local patch based face hallucination methods have been proposed [3–5,7,8]. These approaches assume that the LR and HR patch manifolds share the same local geometrical structure. Specifically, they all directly apply the constructed weights in the LR patch space to the corresponding HR patches. For example, Chang et al. [3] propose a neighbor embedding (NE) image super-resolution method to estimate a HR patch by linearly combining K candidate HR patches with the weights calculated in the LR patch space. For a highly structured object, the position prior of face plays an important role in face image reconstruction and analysis [14–16]. Ma et al. [4] propose a position-patch based face hallucination method through solving a least squares regression (LSR) problem. Most recently, in order to obtain a much more stable and accurate solution, Jung et al. [5] propose to apply the sparsity prior to the patch representation and present a sparse representation (SR) based face hallucination method. In short, the objectives of these local patch based methods [3–5] are try to obtain the best patch representation in the LR patch space, and then transform it to the corresponding HR patches to construct the target HR patch.

However, all of the above methods based on manifold learning can provide good results only with the condition that the image degradation process is simple (e.g., down-sampling by a factor of 2 or 4). However, in many surveillance video applications, there are a variety of factors (such as under-exposure, optical blurring, defocusing, noise and so on [17]) that lead to a significant quality reduction, giving rise to
2.2. The LR layer

In the LR layer, we try to preserve the geometrical structure of the LR patch manifold for the reconstructed HR patch manifold. Following [3], which states that the HR image patches and LR ones form manifolds of the same geometrical structure, we propose in this paper that the LR layer of the neighbor embedding model is characterized by a weight vector that reconstruct one LR input patch from its neighbors in the LR patch manifold formed by all the LR training patches of the same position \((i, j)\), for the sake of convenience we drop the position index \((i, j)\).

\[
J(w, x_H) = \left\| x_L - \sum_{k \in C(x_H)} w_k y_L^k \right\|_2^2 \quad \text{s.t.} \quad \sum_{k=1}^{K} w_k = 1 \quad (1)
\]

where \(C(x_H)\) is the index set of the \(K\)-NN of the estimated HR patch \(x_H\) in the HR training patches \(\{y_H^m(i, j)\}_{m=1}^M\). Different from the traditional manifold learning based super-resolution methods [3, 19], we search the \(K\)-NN in the HR patch manifold by the estimated HR patch.

In addition, we’ve noted that the definition of the weight vector only considers one manifold (the LR patch manifold) and neglects the original HR patch manifold, which is much more credible and discriminant [9, 20]. Therefore, we additionally impose the HR patch reconstruction term to the objective function:

\[
J(w, x_H) = \left\| x_L - \sum_{k \in C(x_H)} w_k y_L^k \right\|_2^2 + \alpha \left\| x_H - \sum_{k \in C(x_H)} w_k y_H^k \right\|_2^2 \quad \text{s.t.} \quad \sum_{k=1}^{K} w_k = 1 \quad (2)
\]

where \(\alpha\) is a parameter balancing the contribution of these two terms and the index \(k\) is in the set \(C(x_H)\) unless otherwise stated. Note that \(x_H\) is the target HR patch and we don’t know it in advance. The objection function (2) with respect to \(w\) and \(x_H\) can be solved respectively and iteratively.

2.3. The HR layer

In the HR layer, we try to apply the geometrical structure of the original HR patch manifold to regularize the reconstructed HR patch manifold. In particular, from a view of graph-based learning, for any two HR patches in the \(K\)-NN set \(\{y_H^k\}_{k \in C'(x_H)}\), the smaller their local distance, the more likely they should have the same weight [15, 16].

Now we describe how to explore the intrinsic geometry (similarity) in the data matrix \(\{y_H^k\}_{k \in C'(x_H)}\) with size of \(K \times K\). Similar with the work of [16], the penalty weighting matrix \(S\) with size of \(K \times K\) is defined on the neighborhoods of the data points as follows:

\[
s_{ij} = \begin{cases} 
\exp \left(-\frac{\|y_H^i - y_H^j\|_2^2}{\sigma^2}\right) & i, j = 1, ..., K \\
0 & \text{otherwise}
\end{cases} \quad (3)
\]

Here, \(\sigma\) is the parameter specifying the width of Gaussian function, which is set to \(\tau \sum_{i,j} \| y_H^i - y_H^j \|_2^2 / K^2\), where \(\tau\)
is a scale parameter. The objective function with our choice of symmetric weights $S$ incurs a heavy penalty if the weights of neighboring points $y_{iL}$ and $y_{jL}$ are very different. Therefore, minimizing it is an attempt to ensure that, if $y_{iL}$ and $y_{jL}$ are “close”, then $w_i$ and $w_j$ are close as well. Following some simple algebraic steps, $\text{tr}(AB) = \text{tr}(BA)$ and $\text{tr}(A) = \text{tr}(A^T)$, we see that

$$\frac{1}{2} \sum_{i,j=1}^{K} (w_i - w_j)^2 S_{i,j} = \frac{1}{2} \sum_{i,j=1}^{K} w_i S_{i,j} w_j^T - \sum_{i,j=1}^{K} w_i S_{i,j} w_j \quad (4)$$

$$= \sum_{i=1}^{K} w_i D_{i} w_i^T - w^T S w^T$$

$$= w(D - S)w^T$$

$$= wLw^T$$

where $D$ is a diagonal matrix, and its entries are column (or row since $S$ is symmetric) sums of $S$. $D_{ii} = \sum_j s_{ij}$, $L = D - S$ is the Laplacian matrix [21]. Here, the superscript "T" means transpose.

### 2.4. Objective Function and Optimization

Considering both of the two properties we want to engage: 1, LR patch manifold neighbor embedding; 2, HR patch manifold preservation on the reconstructed HR patch manifold, the objective function of our proposed CLNE model is defined as:

$$J(w, z_H) = \left\| x_L - \sum_{k} w_k y_{H}^{k} \right\|^2_2 + \alpha \left\| x_H - \sum_{k} w_k y_{H}^{k} \right\|^2_2 + \beta wLw^T \quad (5)$$

where $\beta$ is a parameter controls the tradeoff between matching the LR input patch and finding a HR patch that is compatible with its neighbors. The constrained optimization (5) can be reformulated as

$$J(w, z_x) = \left\| z_x - \sum_{k} w_k y_{x}^{k} \right\|^2_2 + \beta wLw^T \quad (6)$$

where $z_x = \left( \frac{x_L}{\sqrt{\alpha} x_H} \right)$ and $y_{x}^{k} = \left( \frac{y_{H}^{k}}{\sqrt{\alpha} y_{H}^{k}} \right)$.

The minimization of $J(w, z_x)$ with respect to $w$ and $z_x$ can be solved respectively and iteratively. For the $p$-th block, firstly set $z_x(p) = \sum_k w_k y_{x}^{k}$ with the initialized $w$. Then update the weights $w$ with $x_L$ by minimizing $J(w, Z_x)$ as a constrained least squares problem [3, 22]. Note that the optimal weight vector has a closed-form solution given by:

$$w = (G + \beta L)^{-1} \quad (7)$$

where $G$ is the local covariance matrix for $z_x$ as:

$$G = CC^T$$

Define $C$ as:

$$C = (z_x \cdot \text{ones}(1, K) - Z_x^K)$$

where $Z_x^K$ is a matrix with its columns being $z_{x}^{k}$, $\text{ones}(1, K)$ is a $1 \times K$ row vector of ones. The final optimal weight is obtained by rescaling it so that $\sum w_k = 1$.

---

### Algorithm 1 Face Hallucination via CLNE.

**Input:**
- LR and HR training sets $\{Y_L^{m}\}_{m=1}^{M}$ and $\{Y_{H}^{m}\}_{m=1}^{M}$, and a LR face image $X_L$. The parameters: $\alpha$, $\beta$, $K$, and maxIter.

**Output:**
- HR hallucinated face image $X_H$.

1: **Initialize:** $p = 0$, $X_{H}(0) = \text{Bicubic}(X_{L})$.
2: for each LR patch $X_{L}(i, j)$ of $X_{L}$ do
3: repeat
4: Find the $K$-NN index set $C(x_{H})$ and the corresponding $K$-NN HR patch set of $x_{H}$.
5: Calculate the similarity matrix $S$ for the $K$-NN HR patch set of $x_{H}$.
6: Calculate the Laplacian matrix $L = D - S$, $D_{ii} = \sum_j s_{ij}$.
7: Obtain the optimal weights through Eq. (7)-(9).
8: $x_{H} = \sum_{k \in C(x_{H})} w_k y_{x}^{k}$.
9: until $p > \text{maxIter} - 1$
10: end for
11: Integrate all the obtained HR patches above according to the position. The final HR image $X_{H}$ can be generated by averaging pixel values in the overlapping regions.

---

### 3. EXPERIMENTAL VALIDATION

#### 3.1. Datasets and Parameters Settings

The experiment is conducted on FEI face dataset [23], which contains 400 images from 200 subjects and each subject has two frontal images, one with a neutral expression and the other with a smiling facial expression. All the images are cropped to $144 \times 120$ pixels, and we randomly choose 360 images (180 subjects) as the training set, leaving the rest 40 images (20 subjects) for testing. In order to simulate the surveillance, the LR images are obtained by very strong smooth and large down-sampling factor. Specially, the LR images are formed by smoothing (an averaging filter of size $8 \times 8$), down-sampling (by a factor of 8, thus the size of LR face images are $18 \times 15$ pixels), adding Poisson noise (which can better models real noise) to the corresponding HR images. Note that we first upsample the LR image (by factor of 2) to $36 \times 30$ pixels using Bicubic interpolation, and then extract the image pixels as the feature vector according to top-to-bottom and left-to-right scanning to form the feature vector for each patch.
For these local patch based methods, we recommend to use the size of $20 \times 20$ pixels for HR patch and the overlap between neighbor patches is 16 pixels, while corresponding LR patch size is set to $5 \times 5$ pixels with a overlap of 4 pixels. As for the proposed method, the parameters are tuned carefully: $\alpha = 0.01$, $\beta = 0.2$, $K = 75$, and the max iteration $\text{maxIter}$ is set to 5 (the convergence can be achieved very fast).

3.2. Comparison Results

Several typical face hallucination techniques: Bicubic interpolation, Chang et al.’s neighbor embedding (NE) method [3], Ma et al.’s least squares representation (LSR) method [4] and Jung et al.’s sparse representation (SR) method [5] are taken for comparison with the proposed approach. The parameters of all comparison methods are set to the best performance. We compare the above face hallucination approaches quantitatively in terms of the PSNR and SSIM index [24] in Fig. 2. We can see that our approach gives the most faithful hallucination results to the original images. Specifically, the average PSNR and SSIM improvements of CLNE based face hallucination method over the second best method (i.e., Jung’s SR based method [5]) are 0.75 dB and 0.0518, respectively.

Some visual results are shown in Fig. 3. It can be observed that the hallucinated results of Bicubic interpolation are with blurred and jagged artifacts along face counters. Ma’s LSR method [4] will add the noise rather than remove it. This result is also consistent with the objective evaluation from Fig. 2 (LSR has significantly lower SSIM value than other comparison methods). The main reason for this is attributed to its unstable solution. NE [3] and SR [5] can remove noise to some extent. However, the hallucinated faces of NE [3] are blurred and lack of facial details (Fig. 3 (c)), while results of SR [5] are dirty (Fig. 3 (e)). Our proposed method (Fig. 3 (f)) outperforms all the comparison methods and generates the reasonable results with much more facial details.

4. MAIN FINDINGS AND FUTURE DIRECTIONS

The coupled-layer neighbor embedding (CLNE) has been proposed as a novel neighbor embedding model for face image hallucination. In CLNE, we first use the Bicubic interpolator to obtain the initial HR patch, then search the $K$-NN in the HR patch space and calculate the similarity graph in the $K$-NN, and then obtain the reconstruction weights by minimizing the reconstruction error and preserving the graph for the constructed HR patch manifold. After several iteration steps, we can get the target HR patch image. Concatenating and integrating all the hallucinated HR patches, we generate the target HR face image. Experimental results demonstrate the effectiveness of the proposed approach.

5. ACKNOWLEDGMENT

The research was supported by the major national science and technology special projects (2010ZX03004-003-03, 2010ZX03004-001-03), the National Basic Research Program of China (973 Program) (2009CB320906), the National Natural Science Foundation of China (61231015, 61172173, 61070080, 61003184, 61170023), the science and technology Program of Wuhan (201271031366).
6. REFERENCES


