Layered Local Prediction Network with Dynamic Learning for Face Super-resolution

Dahua Lin
Dept. of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong SAR
Email: dhlin4@ie.cuhk.edu.hk

Wei Liu
Dept. of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong SAR
Email: wliu@ie.cuhk.edu.hk

Xiaoou Tang
Dept. of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong SAR
Email: xitang@microsoft.com

***Abstract*— In this paper, we propose a novel framework for face super-resolution based on a layered predictor network. In the first layer, multiple predictors are trained online with a dynamic-constructed training set, which is adaptively selected in order to make the trained model tailored to the testing face. When the dynamic training set is obtained, the optimum predictor can be learned based on the Resampling-Maximum Likelihood-Model. To further enhance the robustness of prediction and the smoothness of the hallucinated image, additional layers are designed to fuse multiple predictors with the fusion rule learned from the training set. Experiments fully demonstrate the effectiveness of the framework.**

I. Introduction

Face super-resolution aims at recovering the high-resolution image from a corresponding low-resolution image, which has received increasing attention in both computer vision and computer graphics.

Recently, a number of algorithms for face super-resolution have been proposed [1] [2] [3] [4]. Baker. et al [1] [2] proposes the Gradient Prior Prediction algorithm with an MAP framework incorporated. However, the prediction of each pixel is merely based on the information within that position, which often leads to instable and noisy results. Liu et al. [3] develops a framework integrating a global parametric linear model and a local patch-based Markov Random Field. Though the spatial information is fully utilized, some person-special details cannot be captured in the trained model. Motivated by Locally Linear Embedding(LLE), Chang et al. [4] models the relation between the low-resolution patches and the high-resolution counter parts by local geometric invariance in the sample space, however the spatial correlation is not sufficiently utilized with a simple patch-decomposition scheme.

A novel framework is presented in this paper in order to solve the following problems: (1) Recovering the detailed features by effectively utilizing both spatial relation and intra-sample relation; (2) Preserving the intra-pixel continuity (smoothness); (3) Reinforcing the robustness and stability of the prediction. To address the first two problems, we develop a scheme, where local predictors which make prediction based on a neighborhood are learned for each pixel respectively. The main novelty is that we propose a dynamic learning process where training sets are adaptively constructed based on the characteristics of the testing face with spatial relation and intra-sample relation taken into account. This process is theoretically justified by the Resampling-Maximum Likelihood Formulation. In addition, the continuous transition of training set across adjacent pixels' predictors ensure the continuity of predictor coefficients and thus the image smoothness. To solve the last problem, we connect all predictors to form a network and introduce additional layers to fuse the predictors with the fusion rule learned from training samples. Experiment results demonstrate the effectiveness of our framework.

II. Dynamic Learning of Local Predictor

A. Motivation of Dynamic Learning

For convenience of describing the neighborhood, we decompose the process of super-resolution into a two-stage procedure: first, we generate an initial super-resolution image of the target size by simple bilinear interpolation, which is named reference image; then an high-quality super-resolution image of the same size is inferred from the reference image. We denote the pixel value at position x on the reference image as $v^L(x)$, and that on the target high-resolution image as $v^H(x)$.

Generally speaking, there exist high correlations between neighboring pixels in a face image, called intra-pixel relation. Based on which, any pixel can be predicted according to its neighboring pixels with high accuracy. This can be formulated with Markov Random Field (MRF).

However, we should note the following two facts: (1) The statistical characteristics of the face image is not the same all over the image, thus the MRF for modelling the face image should be considered as non-stationary. (2) Even in the same position, the local dependencies may change for different face types. Hence, the prediction model should be variant with different positions or different region types.

Traditional interpolation schemes apply fixed prediction filter to all images and all positions in spite of the diverisity in the statistical structure, which tends to oversimplify the situation and fails to recover detailed information. Patch-based hallucination algorithms [3] have successfully exploited the spatial-varying trait and produced results of much higher quality. However, some personal specialities are still lost

0-7803-9134-9/05/$20.00

without exploration into the characteristics of different types of persons.

However, though there exist disparities between different pixels and different samples, it is reasonable to assume that the near samples have similar statistical structure. The assumption constitutes the foundation of adaptive learning.

B. Principle of Dynamic Learning

Motivated from the above observations and analysis, we develop our framework based on a fundamental principle - local structure similarity. Here, "local" refers to not only the spatial relation on the image plane but also the sample relation consisting in the sample space. The principle states that the neighborhood at near positions and in similar regions share nearly the same local dependency structure. But the similarity cannot be propagated to the irrelevant neighborhoods. Accordingly, the predictors should be learned from only the relevant pixels and their neighborhoods with the relevance determined by both spatial distance and similarity of the region that the pixels belong to. To obtain more optimal predictors, the contribution of each training sample should vary with relevances.

Intuitively speaking, the selection of training set based on the relevance-principle has two advantages: first the predictor trained on an adaptively selected set can better model the special local structure without the irrelevant interference. secondly it will result in smooth transition of the training sample weights for adjacent predictors, thus guarantee the continuity of the generated images.

Besides, the measurement of the region similarity deserves special attention. Considering the relative independence of different components of a face such as eyes and mouth, the face is divided into components and the evaluation is done in a component-based manner.

C. Dynamic Construction Algorithm

The local linear predictors are constructed as follows:

The reference image is decomposed into R regions, such as eyes and mouth, denoted as $\mathcal{R}^1, \mathcal{R}^2, \ldots, \mathcal{R}^R$ with pixel values rearranged into R region vectors: $\mathbf{r}_i^1, \mathbf{r}_i^2, \ldots, \mathbf{r}_i^R$. here i is the index of training sample.

When a testing low-resolution face image comes, the reference image is interpolated at first, denoted as I^L. Then the following steps are repeated for each region $\mathcal{R}^r$.

1) Extract the region vector $\mathbf{r}^r$, and find the K nearest vectors: $\mathbf{r}_{i_1}^r, \mathbf{r}_{i_2}^r, \ldots, \mathbf{r}_{i_K}^r$ from the training region vectors of the same region in terms of Euclidean distance between the r-th region vector of testing face and that of i-th training face, denoted as $d_{reg}(r,i) = ||\mathbf{r}^r - \mathbf{r}_i^r||$. The indices $(i_1, r), (i_2, r), \ldots, (i_K, r)$ are also recorded.
2) For each position p in $\mathcal{R}^r$, the following steps are repeated:
   a) Denote the positions in p's neighborhood as $p_1, p_2, \ldots, p_n$. Then we have Kn training samples at n positions on K different faces. For these samples, we extract the neighborhood vectors as $\mathbf{v}_{i_1}^L(p_1), \ldots, \mathbf{v}_{i_1}^L(p_n), \ldots, \mathbf{v}_{i_K}^L(p_1), \ldots, \mathbf{v}_{i_K}^L(p_n)$. and the corresponding target values as $v_{i_1}^H(p_1), \ldots, v_{i_1}^H(p_n), \ldots, v_{i_K}^H(p_1), \ldots, v_{i_K}^H(p_n)$.
   b) Then the prediction model at the position p can be trained based on Kn samples with the Resampling-Maximum Likelihood Formulation as discussed below.

D. Resampling-Maximum Likelihood-Formulation

In our framework, we employ a simple yet effective linear prediction scheme, where the high resolution values are modelled by conditional Gaussian distribution.

$$p\left(v^H(p)|\mathbf{v}^L(p)\right) \propto \exp\left(-\frac{[v^H(p) - \mathrm{pr}(\mathbf{v}^L(p))]^2}{2\sigma^2}\right), \quad (1)$$

$$\mathrm{pr}(\mathbf{v}^L(p)) = \mathbf{c}^T\mathbf{v}^L(p). \quad (2)$$

To establish the relation of dynamic learning and the traditional classification problem, we introduce a concept *Local Structure Class*, denoted as C, which is composed of neighbor vectors extracted from the neighborhoods whose local dependencies are characterized by a same set of coefficients $\mathbf{c}$. With Gaussian distribution assumption, the probability that a neighbor vector at position q from i-th training face belongs to C is

$$p(\mathbf{v}_i^L(q)|C) \propto \exp\left(-\frac{1}{2}\left[\frac{d_{reg}^2(r,i)}{\sigma_{reg}^2} + \frac{d_{pos}^2(r,q)}{\sigma_{pos}^2}\right]\right), \quad (3)$$

where C denotes the local structure class being constructed, $d_{pos}(r,q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$ is the spatial distance. Approximately, the samples other than the Kn selected ones are regarded to have a zero probability.

In fact, the sample are actually not randomly sampled. To simulate the random sampling, we can generate more samples by repeating each selected samples for $N_0 p(\mathbf{v}_i^L(q)|C)$ times, then the training can be based on expanded training set. We will see that the concrete value of N_0 do not influence the result, we use it just for introducing the following deduction.

Considering the expanded training set for class C where the samples can be seen as randomly sampled, the joint likelihood of coefficients $\mathbf{c}$ can be expressed as

$$L(\mathbf{c}, C) = \prod_{k=1}^{K}\prod_{j=1}^{n}\left[p\left(v_{i_k}^H(p_j)|\mathbf{v}_{i_k}^L(p_j)\right)\right]^{N_0 p(\mathbf{v}_{i_k}^L(p_j)|C)}. \quad (4)$$

From (1), (2) and (3), we have

$$-\log L(\mathbf{c}, C) \propto \sum_{k=1}^{K}\sum_{j=1}^{n}\exp\left(-\frac{1}{2}d^2(i_k, j)\right)\epsilon^2(i_k, j), \quad (5)$$

where

$$d^2(i_k, j) = \frac{d_{reg}^2(i_k, p_j)}{\sigma_{reg}^2} + \frac{d_{pos}^2(r, p_j)}{\sigma_{pos}^2}, \quad (6)$$

$$\epsilon^2(i_k, j) = \left[v_{i_k}^H(p_j) - \mathbf{c}^T \mathbf{v}_{i_k}^L(p_j)\right]^2. \quad (7)$$

Then the coefficients $\mathbf{c}$ for the local structure class C can be solved by maximizing the likelihood

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} L(\mathbf{c}, C) = \arg\min_{\mathbf{c}}(-\log L(\mathbf{c}, C)). \quad (8)$$

Denote $\lambda(i, j) = \exp\left(-\frac{1}{2}d^2(i, j)\right)$, which is independent of $\mathbf{c}$, then the problem is converted to a weighted least square optimization problem

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \sum_{k=1}^{K}\sum_{j=1}^{n} \lambda(i_k, j)\left[v_{i_k}^H(p_j) - \mathbf{c}^T \mathbf{v}_{i_k}^L(p_j)\right]^2. \quad (9)$$

It can be easily obtained that the solution to (9) is

$$\hat{\mathbf{c}} = (\mathbf{V}_L \mathbf{\Lambda} \mathbf{V}_L^T)^{-1}(\mathbf{V}_L \mathbf{\Lambda} \mathbf{v}_H). \quad (10)$$

Here $\mathbf{V}^L$ is the $n \times Kn$ matrix:
$\mathbf{V}^L = \left[\mathbf{v}_{i_1}^L(p_1), \ldots, \mathbf{v}_{i_1}^L(p_n), \ldots, \mathbf{v}_{i_K}^L(p_1), \ldots, \mathbf{v}_{i_K}^L(p_n)\right]$
$\mathbf{v}^H$ is a $Kn \times 1$ vector:
$\mathbf{v}^H = \left[v_{i_1}^H(p_1), \ldots, v_{i_1}^H(p_n), \ldots, v_{i_K}^H(p_1), \ldots, v_{i_K}^H(p_n)\right]^T$
$\mathbf{\Lambda}$ is a $Kn \times Kn$ diagonal matrix:
$\mathbf{\Lambda} = \mathrm{diag}\left(\lambda(i_1, 1), \ldots, \lambda(i_1, n), \ldots, \lambda(i_K, 1), \ldots, \lambda(i_K, n)\right)$

Re-sampling is a virtual process reflecting a mechanism to adjust the contributions of different training samples, the coefficients can be calculated by (10).

III. Network-based Fusion

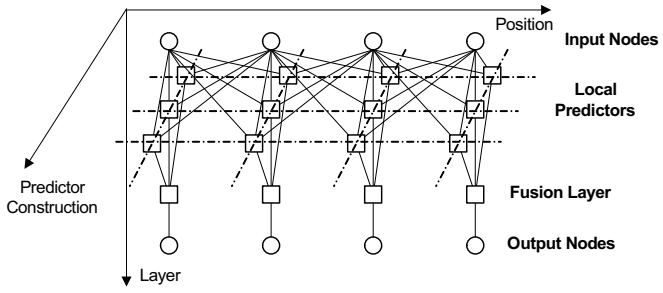

Fig. 2. Layered Local Predictor Network Architecture

A. Fusion of Predictors

To tackle the limited capability and the instability of a single linear predictor, we adopt the fusion strategy to combine multiple predictors into a more robust predictor.

By employing different region-similarity measurements such as Pixel Value-based Euclidean distance, Gradient Field-based Euclidean distance; and different K values, multiple predictors for every pixel are trained. They are combined by weighted sum on the prediction results. Suppose that there are M different predictors generated for each position, the final prediction is expressed as

$$\hat{v}^H = \sum_{m=1}^{M} w_m \mathrm{pr}(\mathbf{v}^L) = \sum_{m=1}^{M} w_m \mathbf{c}_m^T \mathbf{v}^L = \mathbf{w}^T \mathbf{C}^T \mathbf{v}^L. \quad (11)$$

Here, $\mathbf{w} = [w_1, w_2, \ldots, w_M]^T$ is the weighting functions, satisfying $\sum_{m=1}^{M} w_m = 1$. $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M]$ is the prediction coefficients of the predictors. The weights can be obtained by solving a constrained optimization problem:

$$\mathbf{w} = \arg\max_{\mathbf{w}^T\mathbf{1}=1} ||(\mathbf{v}^H)^T - \mathbf{w}^T\mathbf{M}||^2, \quad (12)$$

$$= (\mathbf{1}^T\mathbf{S}^{-1}\mathbf{1})^{-1}\mathbf{S}^{-1}\mathbf{1}, \quad (13)$$

where

$$\mathbf{S} = (\mathbf{v}^H\mathbf{1}^T - \mathbf{M}^T)^T(\mathbf{v}^H\mathbf{1}^T - \mathbf{M}^T) \quad (14)$$

B. Network Architecture

The whole Layered Local Predictors Network (LLPNet) integrates the the local predictors and fusion strategy into a two-layered infrastructure as illustrated as Fig.2.

IV. Experiments

We conduct experiments on a set of frontal face images selected from FERET [5] and XM2VTS [6]. The training set consists of 1360 high resolution images, which are preprocessed with geometric normalization by fixing the positions of eyes and mouth centers and cropped to the size of 96×128. The corresponding low resolution images are obtained by low-pass filtering and down-sampling to size of 24×32.

For testing our framework, experiments are done as follows: all training samples are grouped to an ensemble. When an testing low resolution image is input, an high resolution image is hallucinated employing the procedures described in previous sections. In the dynamic training process, we select $K = 3$ as the number of nearest regions and 3×3 as neighboring window size. Then according to our algorithm, each local predictor has 9 coefficients, which are trained on 27 relevant samples with different weights. Though we need to train 12288 local predictors for an image, however, the computation amount for training each predictor is very small, each image can be produced in 10 - 20 seconds in a Pentium-4 machine. For comparison, other algorithms including B-Spline interpolation and Baker's Method are also implemented and tested.

Comparative results are illustrated in Fig.1. The B-Spline interpolation overblurs the images by using a set of fixed coefficients. Compared to traditional interpolation, Baker's method does clarify some details; however, it incurs notable distortion and noise, in addition, some important detailed features are still missing. By effectively utilizing spatial correlation and intra-sample relation, the local predictors yield significantly better results. However, there still exists a small amount of noise in the results, by further fusing predictors with the learned fusion rule, we can obtain the super-resolution images of much better quality.

V. Conclusion

By integrating the dynamic-constructed local predictors and learning-based fusion strategy into a network, our algorithm achieves encouraging hallucinated results with both person-special details and image smoothness well preserved.

Fig. 1. Results: Column 1: Low Resolution Image, Column 2: Cubic-B Spline Interpolation, Column 3: Baker's Method, Column 4: Results obtained by Local Predictors without Fusion, Column 5: Fusing Predictors by Layered Local Predictor Network, Column 6: Original High-Resolution Image

ACKNOWLEDGEMENTS

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong SAR. The work was done while all the authors are with the Chinese University of Hong Kong.

REFERENCES

[1] S. Baker and T. Kanade, "Hallucinating faces," *Proceedings of ICAFGR*, 2000.

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

[3] C.Liu, H.Shum, and C.Zhang, "A two-step approach to hallucinating faces global parametric model and local nonparametric model," *Proceedings of CVPR01*, 2001.

[4] H. Chang, D. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," *Proceedings of CVPR04*, 2004.

[5] P. J. Phillips, H. Moon, S. A. Ryzvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Trans. PAMI*, vol. 22, no. 10, pp. 1090–1104, 2000.

[6] J. Luettin and G. Maitre, "Evaluation protocol for the extended m2vts database (xm2vts).," *DMI for Perceptual Perceptual Artificial Intelligence*, 1998.