ZOOM BASED SUPER-RESOLUTION THROUGH SAR MODEL FITTING

M. V. Joshi and Subhasis Chaudhuri

Department of Electrical Engineering, Indian Institute of Technology - Bombay
Mumbai 400076. India
{mvjoshi,sc}@ee.iitb.ac.in

ABSTRACT

We propose a technique for super-resolution imaging of a scene from observations at different camera zooms. Given a sequence of images with different zoom factors of a static scene, we obtain a picture of the entire scene at a resolution corresponding to the most zoomed observation. We model the high resolution image as a simultaneous autoregressive (SAR) model, the parameters of which are learnt from the most zoomed observation. Assuming that the entire scene can be described by a homogeneous SAR model, the learnt parameters are then used in a suitable regularization technique to estimate the high resolution field.

1. INTRODUCTION

Super-resolution refers to the process of producing a high spatial resolution image from several low-resolution observations. It includes upsampling the image thereby increasing the maximum spatial frequency and removing degradations that arise during image capture, viz., aliasing and blurring. The amount of aliasing differs with zooming. The larger the scene coverage, the lower will be the resolution with more aliasing effect. Thus one can use zoom as a cue for generating high-resolution images at the lesser zoomed area of a scene. Immersive viewing on the Internet is one such application where the least zoomed entire scene or a portion of it can be viewed at a higher resolution by using the zoomed observations.

Many researchers have tackled the super-resolution problem for both still and video images, e.g., [6, 10, 9] (see [4] for details). Recently, the researchers are exploiting the use of learning based techniques for super-resolution. Capel and Zisserman have proposed a super-resolution technique from multiple views using learnt image models [3]. Their method uses learnt image models either to directly constrain the ML estimate or as a prior for a MAP estimate. Authors in [5] describe image interpolation algorithms which use a database of training images to create plausible high frequency details in zoomed images. In [1] authors develop a super-resolution algorithm by modifying the prior term in the cost to include the results of a set of recognition decisions, and call it as recognition based super-resolution or hallucination. The prior term enforces the condition that the gradient of the super-resolved image should be equal to the gradient of the best-matched training image. A learning based super-resolution enhancement of video is proposed by Bishop *et al.* [2]. Their approach builds on the principle of example based super-resolution for still images proposed by Freeman *et al.*[5]. They use a learned data set of image patches capturing the relationship between the middle and the high spatial frequency bands of natural images and use an appropriate prior distribution over such patches. The method described here can also be classified under this category of learning based methods, but with a different type of cue for parameter learning.

We now discuss some of the previous works carried out on simultaneous autoregressive (SAR) models for image processing. Kashyap and Chellappa [7] estimate the unknown parameters for the SAR and the conditional Markov (CM) models and also discuss the decision rule for the choice of neighbors using synthetic patterns. Authors in [8] use a multiresolution simultaneous autoregressive model for the texture classification and the segmentation. They derive a rotation invariant SAR model for the texture classification. As discussed in [5], the richness of the real world images would be difficult to capture analytically. This motivates us to use a learning based approach, where the SAR parameters of the super-resolved image can be learnt from the most zoomed observation and hence can be used to estimate the super-resolution image for the least zoomed entire scene. The basic problem that we address in this paper can be defined as follows: One continuously zooms in to a scene while capturing its images. The most zoomed-in observation has the highest spatial resolution. We are interested in generating an image of the entire scene (as observed by the wide angle or the least zoomed view) at the same resolution as the most zoomed-in observation. We model the high resolution image as a homogeneous simultaneous autoregressive (SAR) model, learn the corresponding field parameters from the high resolution observation and use this prior to super-resolve the rest of the scene captured at a lower reso-

0-7803-8554-3/04/$20.00 ©2004 IEEE.

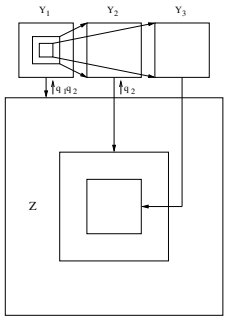


Fig. 1. Illustration of observations at different zoom levels: Y_1 corresponds to the least zoomed and Y_3 to the most zoomed images. Here Z is the high-resolution image.

lution.

2. OBSERVATION MODEL

The zooming based super-resolution problem can be cast in a restoration framework. There are p observed images $\{Y_i\}_{i=1}^{p}$ each captured with different zoom settings and are of size $M \times M$ pixels each. Figure 1 illustrates the block schematic of how the low-resolution observations of a scene at different zoom settings are related to the high-resolution image. Here we consider that the most zoomed observed image of the scene Y_p ($p = 3$ in the figure) has the highest spatial resolution. We are assuming that there is no rotation about the optical axis between the observed images taken at different zooms. However, we do allow a lateral shift of the optical center as the zooming process may physically shift the camera position by a small amount. Since different zoom settings give rise to different resolutions, the least zoomed scene corresponding to entire scene needs to be upsampled to the size of $N \times N$ pixels, where $N = (q_1, q_2, \cdots, q_{p-1}) \times M$ and $q_1, q_2, \cdots, q_{p-1}$ are the corresponding zoom factors between two successively observed images of the scene $Y_1Y_2, Y_2Y_3, \cdots, Y_{(p-1)}Y_p$, respectively. Given Y_p, the remaining $(p-1)$ observed images are then modeled as decimated and noisy versions of this single high-resolution image of the appropriate region in the scene. The most zoomed observed image will have no decimation. Let $\mathbf{z}$ represent the lexicographically ordered high-resolution image of size $N^2 \times 1$ pixels. If $\mathbf{y}_m$ is the $M^2 \times 1$ lexicographically ordered vector containing pixels from differently zoomed images Y_m, the observed images can be modeled as

$$\mathbf{y}_m = D_m \mathcal{C}_m(\mathbf{z} - z_m) + \mathbf{n}_m, \qquad m = 1, \cdots, p \quad (1)$$

where D is the decimation matrix, size of which depends on the zoom factor. For an integer zoom factor of q, the decimation matrix D consists of q^2 non-zero elements of value $\frac{1}{q^2}$ along each row at appropriate locations. Here $\mathcal{C}_m(\mathbf{z} - z_m)$ is a cropping operator with z_m representing the lateral shift of the optical center during the zooming process. The cropping operator is similar to a characteristic function, that crops out $\lfloor q_1 q_2 \cdots q_{m-1} N \rfloor \times \lfloor q_1 q_2 \cdots q_{m-1} N \rfloor$ pixel area from the high resolution image $\mathbf{z}$ at an appropriate position. p is the number of observations, $\mathbf{n}_m$ is the $M^2 \times 1$ i.i.d noise vector with zero mean and variance σ^2. Our problem now reduces to estimating $\mathbf{z}$ given $\mathbf{y}_m$'s, which is an ill-posed, inverse problem.

3. SUPER-RESOLUTION RESTORATION

3.1. Image Field Modeling

The MRF provides a convenient way of modeling context dependent entities such as pixel intensities, depth of the object and other spatially correlated features. This is achieved through characterizing mutual influence among such entities using conditional probabilities for a given neighborhood. Although the MRF model for prior constitutes a popular statistical model, and captures the contextual dependencies very well, the computational complexities with these models are high as one needs to compute the partition function in order to estimate the true parameters. The computational burden can be reduced by using a scheme such as the maximum pseudolikelihood (MPL) which does not involve the estimation of partition function. But to obtain the global minima we still need to use a stochastic relaxation technique, which is computationally taxing. Also the pseudolikelihood is not a true likelihood except for the trivial case of nil neighborhood.This motivates us to use a different but a suitable prior. We can consider the linear dependency of a pixel in a super-resolved image to its neighbors and represent the same by using simultaneous autoregressive (SAR) model and use this SAR model as the prior. Although this becomes a weaker prior the computation is drastically reduced.

Let $z(s)$ be the gray level value of a pixel at site $s = (i, j)$ in an $N \times N$ lattice, where $(i, j) = 1, 2, \cdots N$. The SAR model for $z(s)$ can then be expressed as [7]

$$z(s) = \sum_{r \in \mathcal{N}_s} \theta(r) z(s + r) + \sqrt{\rho} n(s) \qquad (2)$$

where $\mathcal{N}_s$ is the set of neighbors of pixel at s. $\theta(r),\ r \in \mathcal{N}_s$ and ρ are unknown parameters and $n(.)$ is an independent and identically distributed (i.i.d) noise sequence with zero mean and unit variance . We use a fifth order neighborhood that requires a total of 24 parameters $\theta(i, j)$ as shown in Figure 2. In order to reduce the computations while estimating these parameters we use a symmetric SAR model where $\theta(r) = \theta(-r)$ and estimate only 8 parameters. It may be mentioned here that we are not discussing here the choice of appropriate order for the neighborhood system and the choice of number of parameters for optimal results.

3.2. Parameter Learning

In order to enforce the SAR prior, we must know the values of the model parameters $\boldsymbol{\theta}$. These parameters must be learnt from the given observations themselves. The parameters

(–2,–2)	(–2,–1)	(–2, 0)	(–2, 1)	(–2, 2)
(–1,–2)	(–1,–1)	(–1, 0)	(–1, 1)	(–1, 2)
(0, –2)	(0, –1)	(0, 0)	(0, 1)	(0, 2)
(1, –2)	(1, –1)	(1, 0)	(1, 1)	(1, 2)
(2, –2)	(2, –1)	(2, 0)	(2, 1)	(2, 2)

Fig. 2. The fifth order neighborhood for pixel at location $(0, 0)$.

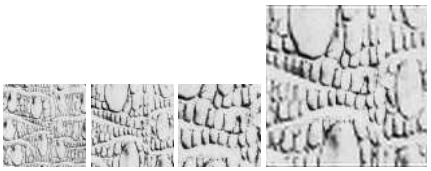
cannot be learnt from the low resolution observations. But, there exists one observation Y_p where a part of the scene is available at the high resolution. Hence, we use the observation Y_p to estimate the parameters. The inherent assumption is that the entire scene is statistically homogeneous and it does not matter which part of the scene is used to learn the model parameters. One of the characteristics of an image data is the statistical dependence of the gray level at a lattice point on those of its neighbors. This statistical dependency can be characterized by using an SAR model where the gray level at a location is expressed as a linear combination of the neighborhood gray levels and an additive noise. We estimate the SAR model parameters by considering the image as a finite lattice model and using the iterative scheme as given in [7], by using the most zoomed image as a SAR model.

3.3. Restoration using SAR Prior

With the SAR parameters estimated we would like to arrive at a cost function which has to be minimized. We use a regularization based approach which is quite amenable to the incorporation of information from multiple observations with the regularization function chosen from the prior knowledge of the function to be estimated. The prior knowledge here, serves as a contextual constraint used to regularize the solution. We use the simple linear dependency of a pixel value on its neighbors as a constraint using the SAR model for the image to be recovered. Here the estimated SAR parameters serve as the coefficients for linear dependency. Using a data fitting term and a prior term one can easily derive the corresponding cost function to be minimized as

$$\begin{aligned} \epsilon &= \lambda \sum_{m=1}^{p} \|\mathbf{y}_m - D_m \mathcal{C}_m(\mathbf{z} - z_m)\|^2 \\ &+ \sum_{i,j} \left(z(s) - \sum_{r \in \mathcal{N}_s} \theta(r) z(s+r) \right)^2 . \end{aligned}$$

Here λ is a regularization parameter which is now proportional to $\frac{\sigma^2}{\rho}$ where ρ is the error variance for the SAR model. Since the model parameter vector $\boldsymbol{\theta}$ has already been estimated, a solution to the above equation is, indeed, possible. The above cost function is convex and is minimized using the gradient descent technique. The initial estimate $\mathbf{z}^{(0)}$ is obtained as follows. Pixels in the bilinearly interpolated least zoomed observed image corresponding to the entire scene is replaced successively at appropriate places with bilinear interpolation of the other observed images with increasing zoom factors. Finally the most zoomed observed image with the highest resolution is copied with no interpolation.



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

Fig. 3. (a-c) Observed images of a texture captured with three different zoom settings. (d) The super-resolved image using the proposed algorithm for a zoom factor of $q = 2$ between (b) and (c).

(a) (b)

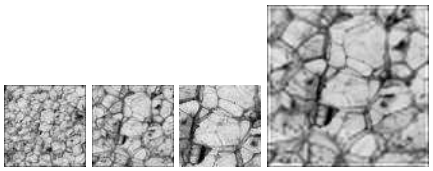
Fig. 4. (a) Zoomed texture image formed by successive bilinear expansion. (b) The super-resolved image for a zoom factor of $q = 4$.

4. RESULTS AND CONCLUSIONS

In this section, we demonstrate the efficacy of the proposed technique to recover the super-resolved image from observations at different zooms through learning of model parameters.

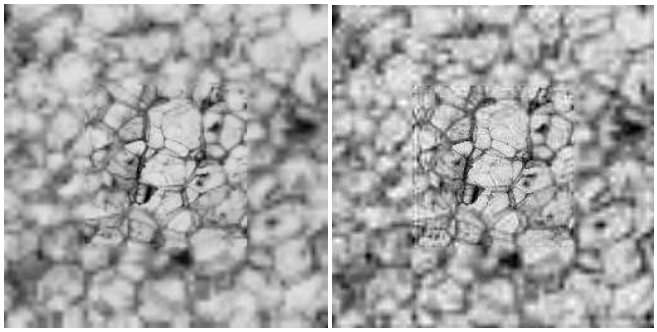
Initially we experimented on simulated data. A number of images were chosen from the Brodatz's album. We observe an image at three levels of zoom $q_1 = q_2 = 2$. Figures 3(a-c) show one such set of observations, where Figure 3(a) shows the entire image at a very low resolution. Figure 3(b) shows one-fourth of the region at double the resolution and Figure 3(c) shows only a small part of Figure 3(a) at the highest resolution.

Using the estimated parameter set, we super-resolve the entire scene in Figure 3(a) to obtain the Figure 4(b). Compare the result to that obtained using a simple bilinear zooming operation given in Figure 4(a). We notice that both the



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

Fig. 5. (a-c) Observed images of another texture captured with three different zoom settings. (d) The super-resolved image for a zoom factor of $q = 2$.



(a) (b)

Fig. 6. (a) Bilinearly zoomed texture image. (b) The super-resolved texture image.

images appear blurred near the periphery. However, the interpolated image is too blurred to infer about the texture. We can easily observe that, the restoration upto a zoom factor $q = 2$ is quite good as is evident from the Figure 3(d). For a zoom factor of $q = 4$, one needs to reconstruct 16 pixels from each observed pixel near the periphery, which is clearly a difficult task. A degradation in the reconstruction is, thus, quite expected even in the estimated high resolution image. The result for another set of observed textures, shown in Figures 5(a-c), is given in Figure 6(b). Once again, a comparison with the corresponding zoomed image in Figure 6(a) brings out a similar conclusion that upto a zoom factor $q = 2$, (see Figure 5(d)) the results of the proposed super-resolution scheme is very good, but beyond that the quality of restoration starts degrading. However the mean squared error comparison for the proposed approach and the successive bilinear interpolated image when measured with respect to the original image showed a significant decrease of about 30% in all of the above experiments.

Next we consider an example of real data captured using a camera with optical zoom. Unlike the experiments on simulated data, the assumption of the homogeneity is not strictly valid for the real data. However in the absence of any other usable priors, we continue to make use of this assumption and we still obtain a reasonably good super-resolution reconstruction. Figures 7(a-c) show the corresponding observations of a house image of size 72×96 each. The zoom factors were carefully chosen such that the relative zoom factors between successive observations are again $q = 2$. Since the images are captured with a varying zoom, there may be a change in AGC of the camera. Hence the observations are mean corrected to alleviate this problem. The experimental results of super-resolution restoration are given in Figures 8(a,b). Similar conclusions can again be drawn from this experiment.

(a) (b) (c)

Fig. 7. (a-c) Observed images of a house captured with three different known zoom settings.

(a) (b)

Fig. 8. (a) Zoomed image formed using successive bilinear expansion, (b) Super-resolved house image using the SAR prior.

We have demonstrated that it is, indeed, possible to obtain a high resolution image of a scene using zoom as a cue and by learning the parameters from the most zoomed observation.

5. REFERENCES

[1] S. Baker and T. Kanade. Limits on Super-Resolution and How to Break Them. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, September 2002.

[2] C. M. Bishop, A. Blake, and B. Marthi. Super-Resolution Enhancement of Video. In *Int. Conf. on Artificial Intelligence and Statistics*, Key West, Florida, 2003.

[3] D. Capel and A. Zisserman. Super-Resolution from Multiple Views using Learnt Image Models. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages II:627–634, 2001.

[4] S. Chaudhuri. *(Ed.), Super-Resolution Imaging*. Kluwer Academic Publisher, Boston, 2001.

[5] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-Based Super-Resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, March/April 2002.

[6] M. Irani and S. Peleg. Improving Resolution by Image Registration. *CVGIP:Graphical Models and Image Processing*, 53:231–239, March 1991.

[7] R. Kashyap and R. Chellappa. Estimation and Choice of Neighbors in Spatial-Interaction Models of Images. *IEEE trans. on Information Theory*, 29(1):60–72, January 1983.

[8] J. Mao and A. K. Jain. Texture Classification and Segmentation using Multiresolution Simultaneous Autoregressive Models. *Pattern Recognition*, 25(2):173–188, 1992.

[9] D. Rajan and S. Chaudhuri. Simultaneous Estimation of Super-Resolved Intensity and Depth Maps from Low Resolution Defocussed Observations of a Scene. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 113–118, Vancouver Canada, 2001.

[10] R. R. Schultz and R. L. Stevenson. Extraction of High-Resolution Frames from Video Sequences. *IEEE Trans. on Image Processing*, 5(6):996–1011, June 1996.