How much zoom is the right zoom from the perspective of Super-Resolution?

Himanshu Arora and Anoop M. Namboodiri
Center for Visual Information Technology, IIIT, Hyderabad, 500032, India
{himanshu@research., anoop@}iiit.ac.in

Abstract

Constructing a high-resolution (HR) image from low-resolution (LR) image(s) has been a very active research topic recently with focus shifting from multi-frames to learning based single-frame super-resolution (SR). Multi-frame SR algorithms attempt the exact reconstruction of reality, but are limited to small magnification factors. Learning based SR algorithms learn the correspondences between LR and HR patches. Accurate replacements or revealing the exact underlying information is not guaranteed in many scenarios. In this paper we propose an alternate solution. We propose to capture images at right zoom such that it has just sufficient amount of information so that further resolution enhancements can be easily achieved using any off the shelf single-frame SR algorithm. This is true under the assumption that such a zoom factor is not very high, which is true for most man-made structures. The low-resolution image is divided into small patches and ideal resolution is predicted for every patch. The contextual information is incorporated using a Markov Random Field based prior. Training data is generated from high-quality images and can use any single-frame SR algorithm. Several constraints are proposed to minimize the extent of zoom-in. We validate the proposed approach on synthetic data and real world images to show the robustness.

1. Introduction

High quality image generation is an important problem that finds various applications in computer vision and image processing. Super-resolution (SR) [20] is the process of generating a high-resolution (HR) image from low-resolution (LR) image(s). Various super-resolution algorithms are commonly divided into two categories, *viz.* multi-frame SR [7] and learning based SR [9]. Lin and Shum [16] showed that the theoretical limit on magnification for multi-frame SR is 5.7, and in practical scenarios this limit is only 2.5. For higher magnification factors, the number of images required increases exponentially, making the computational cost beyond practical limits for most applications. Multi-frame SR also requires accurate registration and blur parameters, which are very difficult to obtain in many scenarios. These drawbacks limit the applicability of multi-frame SR, and it is primarily used for revealing the exact underlying details at a limited magnification. The super-resolved images are useful to achieve higher recognition rates for various vision algorithms, e.g. [2].

In contrast, learning based single image SR, in theory, can achieve magnification factors up to 10, as shown by Lin *et al.* [15]. The HR image generation is formulated as an inference problem. Correspondences between LR and HR patches are stored during the learning phase, and the HR image is inferred in a MRF framework with contextual constraints. This category of algorithms perform well for natural objects, where the perceptual quality is more important than accurate reconstruction of reality. They also work well if the training set is optimized for specific object/scene classes, such as faces [4]. However, the performance drops significantly on man-made structures where even with a magnification factor of 3 (see Fig. 4(a), in [15]), the actual content need not be resolved in the final result.

We note that the bottleneck of a learning based SR algorithm lies in the nature of the underlying data, and the magnification factors achievable for various types of images or regions within an image, vary considerably. In other words, in order to get uniform perceptual quality after SR, different regions of an image need to be captured at different minimum resolutions. One could be conservative, and capture the whole image at the maximum resolution required by any image patch, which is both costly and redundant. Capturing minimum number of images in the whole process require us to use learning based approaches. In this paper, we propose a solution to this problem by capturing the image at ideal resolutions. The minimum required resolution for every patch of the image is predicted from a low-resolution image. Different parts of the image are then captured at the correct resolution, and thus sufficient amount of scene information is gathered at the image capturing stage itself. Any further magnification of the image can be achieved using any off the shelf single image super-resolution algorithm.

978-0-7695-3476-3/08 $25.00 © 2008 IEEE
DOI 10.1109/ICVGIP.2008.57

The ability to predict the ideal resolution for capture of an image region also enables a variety of applications. Automatically selecting the right resolution or zoom would enable efficient mosaicing of very large panoramas. Instead of capturing all the images at a high resolution [12], the final mosaic can be generated with fewer number of images at the right zoom level. The predicted resolution would also represent the minimum amount of information that is essential to represent a scene, and hence would reduce the computational cost of many vision algorithms that attempt scene understanding. Mobile robots could use this information to interpret and navigate the world more efficiently. Removing the redundant information that could be recreated using SR would also enable effective compression.

For most man-made structures, a limit on amount of scene information gathered can be quantified empirically. Note that primitives such as step edges along smooth curves can be enhanced effectively using single-frame SR. On the contrary, for most natural scenes, a very high value of zoom is required because of their detailed and intricate structure. However, one could replace the lost information with high-quality pre-captured content, without affecting the perceptual quality. We formulate the problem of capturing an image at ideal resolution in a patch based framework, where the ideal resolution/zoom is predicted separately for every patch. The ideal resolution or zoom will thus depend on the nature of the scene, the level of detail, and the information that can be captured by learning based SR algorithms, making the prediction challenging. We note the following points about image patches to predict ideal resolution factors.

- The structures in the image are assumed to have edges along smooth curves, which lends to enhancement by SR algorithms. The basic patch provides sufficient information to predict up to *smaller magnifications*.
- For *larger magnification* factors, the context information plays an important role, which is obtained from the predicted zoom values of the neighboring patches.
- The size of the patch is appropriately selected to provide enough structural information for smaller magnification factors and simultaneously include strong context information for predicting larger magnifications.

Once the patch size is selected, we need to learn the prediction function for the zoom level of individual patches, and to model the contextual relationship with neighboring patches. We use a Markov Random Field (MRF) framework, which is popularly used to incorporate contextual constraints.

In short, we propose an approach for high-resolution generation by capturing sufficient information at the image capturing stage itself. The image is decomposed into patches and zoom level prediction is modeled as an inference problem in a MAP-MRF framework. We use Bayesian belief propagation rules to solve the network. As the optimization function contains numerous local minima, a robust technique is proposed to initialize the solution. Various practical constraints are proposed to minimize the extent of zoom-in. The results are validated on synthetic data and experiments are performed on real scenarios as well.

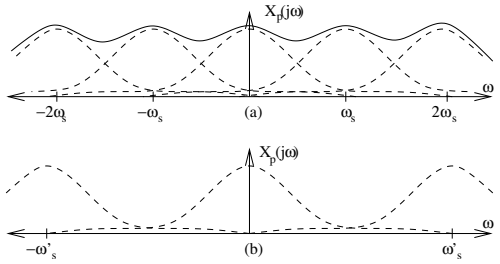


Figure 1. Fourier spectra of a hypothetical signal with different sampling rates; (a) sampling rate is low; (b) sampling rate is high enough so that the image can be zoomed in further easily with minimum aliasing.

2. Related Work

There are different categories of work that address the problem of automated zoom detection from different perspectives. [10, 1] address the problem on zooming in on a pre-determined object by placing it to fill the image or by zooming-in only on the focused areas. Tordoff and Murray [21] model the zoom control for a tracking system. The goal is to zoom-in and out such that the target remains within the field of view of the camera with high confidence.

Image-cropping algorithms [6, 17, 18] can potentially be used to zoom the image to the desired target. The region of potential interest is selected from an image using a predefined criterion. The selected portion can be zoomed in to emulate automatic zooming. However, these algorithms do not address the resolution of the desired object and only directs the attention to it.

In computer vision literature, the term zoom has been used in different contexts. To avoid any ambiguities we mention some of them in related work. Traditionally, zoom-in refers to the change in focal length of the camera lens. Jin *et al.* [11] proposed a probabilistic model to detect zoom-in or zoom-out operations in an image sequence. Zooming-in is also used to refer to magnification of image using super-resolution algorithms, and not by camera e.g. [5].

3. Predicting the Right Zoom

In the paper, the right zoom of the camera is such that the image captured at that zoom contains sufficient amount of information. Image can then be magnified further with simple algorithms which enhances edges and certain features. We first describe 'zooming-in sufficiently' from Nyquist view. Zoom prediction is modeled as an inference problem. The image is divided into patches and zoom factor is

predicted for each patch. Both structural cues and context information around the patch are incorporated and modeled in a MAP-MRF framework. The network is solved using Bayesian belief propagation rules. Randomness measure is defined to initialize zooming factor in the network.

3.1. A Nyquist View of Zoom-in

The irradiance field observed by a camera requires very large frequency range to represent all information. One observation to be made is that the magnitude of Fourier spectra usually decreases as a function of increasing absolute frequency. According to the Nyquist theorem [19], a signal can be uniquely reconstructed from its samples if the size of the band of input signal is less than the sampling frequency. Fig. 1 shows the Fourier spectra of a hypothetical signal at different sampling rates. If sampling rate is low, Fig. 1(a), signal is highly aliased and significant information is lost. On the other hand if sampling rate is high enough, Fig. 1(b), the aliasing is low and significant information can be recovered from the sampled information. Rest of the high frequency are usually step edges, which can be recovered by promoting step functions and edges along smooth curves while zooming-in, and noise, which can be characterized and ignored. This forms the basis of selecting the right zoom of the camera. Sufficient information is gathered at the image capturing stage so that any further resolution enhancements requires only simple feature enhancements.

3.2. Probabilistic Model

Image at the right zoom is captured in two steps. In the first step, a low-resolution image is captured and the zoom is predicted for each patch. In the second step the image(s) are captured at the right zoom. Before we describe probabilistic model we define the *resolution-front* of an image.

Definition 1. *Let $\tilde{I} = \{\tilde{I}_1, \tilde{I}_2, ...\tilde{I}_N\}$ be the image captured, represented as a concatenation of square patches, $\tilde{I}_i$, on a 2D grid each of size $m \times m$ at image locations $1, 2...N$.* **Resolution front** *$R_f = \{f_1, f_2, ...f_N\}$ of the image $\tilde{I}$ is the amount of minimum magnification f_i required at image patch location i, so that the block can be super-resolved further by using only simple feature enhancement algorithms.*

We essentially predict the resolution front rather than the absolute zoom required. It has more usability in various scenarios, some of which are discussed in experiments and results section. The prediction strategy should follow three principles mentioned before. We present our zoom prediction algorithm as an inference problem similar to inference problems presented by Freeman *et al.* [9] in a Markov Network. The Maximum-a-Posteriori (MAP) estimate of the

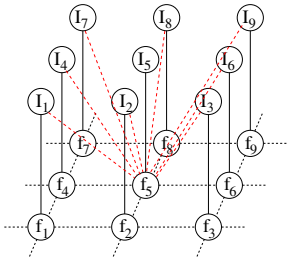


Figure 2. Markov Network for zoom prediction. $\tilde{\mathrm{I}}_i$ are LR patches and the corresponding resolution front values f_i. The output value at any location is also dependent on certain information of neighboring patches and the context.

resolution-front R_{MAP} is given by,

$$\mathrm{R}_{MAP} = \arg\max_{\mathrm{R}_f} P(\mathrm{R}_f|\tilde{\mathrm{I}}), \tag{1}$$

$$= \arg\max_{\mathrm{R}_f} p(\tilde{\mathrm{I}}|\mathrm{R}_f)P(\mathrm{R}_f), \tag{2}$$

To simplify the inference problem, the formulation is reduced to a patch based model under Markov assumption similar to the one used in [9]. However, our patch structure incorporates intensities of all pixels of the underlying patch, $\tilde{\mathrm{I}}_i$, at higher weights and some pixels from neighboring patches at lower weights(see sec. 3.3). Let $\mathbf{y}_i$ be a column vector which contains intensity values of all such pixels. Markov Random Field (MRF) is a popular framework to include contextual constraints. Each node in the network corresponds to either an image patch or a resolution front value. Fig. 2 shows the graphical dependencies among nodes. To maintain compatibilities of resolution-front value predictions with neighbors, a 5-value resolution-front tuple is predicted at each location. It includes the resolution front values corresponding to underlying patch and its 4-neighbors. Let f_i^j denote the resolution front value, predicted using pixel information at patch location i, for patch at location j such that $j \in N(i)$ is one of the 4 neighbors. The maximum likelihood estimate $p(\tilde{\mathrm{I}}|\mathrm{R}_f)$ is,

$$p(\tilde{\mathrm{I}}|\mathrm{R}_f) = \prod_i p(\mathbf{y}_i|f_i, f_i^j)$$

$$= \prod_i \frac{1}{Z} e^{-\frac{1}{2}(\mathbf{x}_i-\mathbf{y}_i)^T\Sigma_1^{-1}(\mathbf{x}_i-\mathbf{y}_i)}, \tag{3}$$

where $\mathbf{x}_i$ is a vector from the training data for which the equation is optimized and the corresponding resolution-front assignment is ML estimate. Σ_1 is a diagonal matrix which incorporates the weights given to different pixel values of the patch. The above equation is also known as pairwise compatibility function between input and output values in a Markov network [9]. The resolution front should be compatible and dependent on the neighboring context,

$$P(\mathrm{R}_f) = \prod_i P(f_i) = \prod_i \prod_{j\in N(i)} P(f_i|f_j). \tag{4}$$

The compatibility function (equivalent to above function

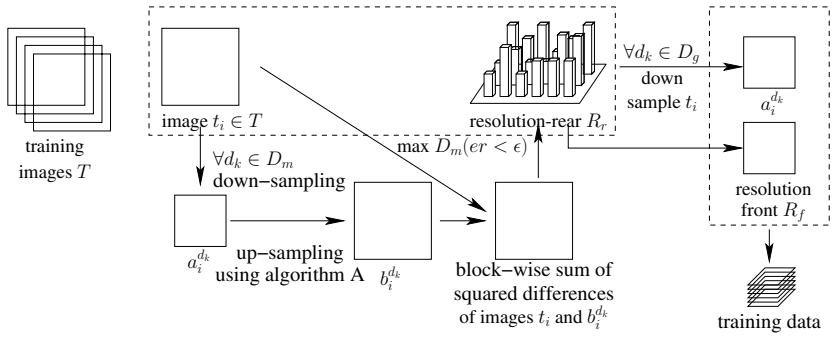


Figure 3. Block diagram explaining the generation of training data.

$P(f_i|f_j)$) between the predicted resolution front values and the neighboring values is proposed as,

$$\psi(f_i, f_j) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\left((f_i - f_j^i)^2 + (f_j - f_i^j)^2\right)/2\sigma_2^2}, \quad (5)$$

where σ_2^2 is the variance. Substituting equation 3 and 5 into equation 1 and after taking the logarithm we get,

$$\mathrm{R}_{MAP} = \underset{\mathrm{R}=\{f_1 \ldots f_N\}}{\arg\min} \left(\sum_i (\mathbf{x}_i - \mathbf{y}_i)^T \Sigma_1^{-1} (\mathbf{x}_i - \mathbf{y}_i) + \sum_i \sum_{j \in N(i)} \left((f_i - f_j^i)^2 + (f_j - f_i^j)^2 \right) / 2\sigma_2^2 \right) \quad (6)$$

3.3. Patch Representation

Smaller patch size is desirable to increase the generalizability and larger patches for specificity. Square patch sizes having equal weights to all pixels are commonly used in a Markov network. We use slightly larger patch sizes but assign low weights to the pixels away from the center patch while computing the L_2 distance. Equation 7 is intuitive and this patch representation is used. This behavior is embedded in Σ_1 in the MAP formulation. The function describing the patch model is,

$$f(\mathbf{x}) = \begin{cases} c, & |\mathbf{x}| <= \mathbf{t} \\ \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp(-\frac{\mathbf{x}^2}{2\sigma_p^2}), & \mathbf{t} < |\mathbf{x}| <= \mathbf{p} \end{cases} \quad (7)$$

where c is a constant and $2t+1$ is the underlying patch size.

3.4. Training Data Generation

Training patches are generated from selected images which a user believes that can be super-resolved further by using any single-frame SR algorithm. To simplify descriptions, we define *resolution-rear* similar to resolution-front as,

Definition 2. *Let $I = \{I_1, I_2, ...I_N\}$ be the given image, represented as a concatenation of square patches, I_i each of size $m \times m$ at image locations $1, 2...N$.* **Resolution rear** *$R_r = \{r_1, r_2, ...r_N\}$ of the image I is the amount of maximum down-sampling, r_i at image patch location i, so that the down-sampled block can be super-resolved to the original block I_i by using only simple feature enhancement algorithms. The image I has a resolution front value $R_f = \{1\}$.*

For each image t_i, from the training images **T**, we calculate how much down-sampling each block can tolerate. We downsample[1] the image at various downsampling values $D_m = \{d_1, d_2, ...d_t\}$ and then super-resolve the image using an algorithm **A**. Block-wise sum of squared differences in intensity values is computed between the original and super-resolved image. If the error is greater than a threshold ϵ then then downsampling factor just smaller than current downsampling factor is assigned to r_i. For any down-sampled (by a factor k)version of image I, the resolution-front is computed from resolution-rear as,

$$f_i = \begin{cases} \frac{k}{r_i}, & \text{if } \left(\frac{k}{r_i}\right) > 1 \\ 1, & \text{otherwise} \end{cases} \quad (8)$$

The original image is downsampled at multiple resolution factors in D_g and resolution front is computed for each of them. 5-value tuple having resolution front value of the patch and its 4-neighbors are stored along with the patch in the training database. Fig. 3 explains the training patch generation process. When an image is down-sampled the block size varies at different downsampling factors. For the sake of efficiency in searching, a constant patch size is required. We take a constant block size and assign the second highest (to avoid outliers) resolution-front value. Training data is generated at various equally spaced non-integer zoom values as well. Fig. 4(a) shows patch intensity structures corresponding to integer zooms only.

For higher accuracy, images at various resolutions should be captured from the camera. We prefer to downsample images offline because, a) Computation of lens distortions parameters, which are different at different focal lengths, and estimation of registration parameters need to be highly accurate and the process is computationally ex-

[1] downsampling factor=1/scaling factor

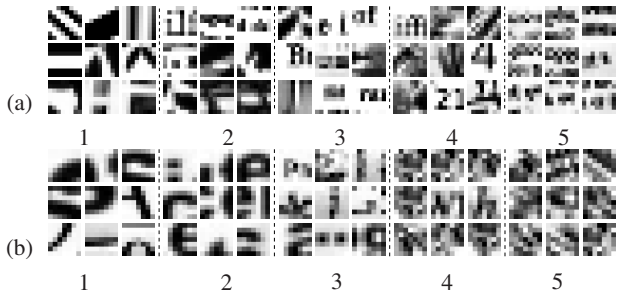


Figure 4. some patch structures and corresponding zoom-in values (a) computed in training phase. 4×4 is the central patch and 8×8 is overall patch with pixels from neighbors. (b) using randomness measure (sec 3.6).

pensive; b) Varying degree of error in measurements irradiance field and presence of noise. Certain relaxations are incorporated in error limits at various stages.

3.5. Energy Minimization

Solving equation 6 for global minima is computationally prohibitive with large number of patches. Freeman *et al.* [9] favored to obtain a local minimal solution which approximates the global minima. Using approximate nearest neighbor data-structure [3] a smaller set of similar patches (usually 20-30) are obtained. Markov network is solved using local message passing algorithm (belief propagation). Rules are same as proposed in [9]. It was argued that these rules can be applied on graphs with loops as well without significant deviation from solution. However, presence of multiple local minimas requires a robust initialization, which is often ignored. In the next subsection, a general method is proposed to initialize the resolution-front values.

3.6. Robust Initialization

Each pixel value is initialized to a zoom value proportional to randomness in intensity structure in the neighborhood. Randomness measure P_i at location i is proposed as,

$$P_i = \sum_{j=[-3,3]\times[-3,3]} \boldsymbol{\sigma}_{gr_3}(i+j)\boldsymbol{\sigma}_{in_3}(i+j), \qquad (9)$$

where $\boldsymbol{\sigma}_{in_3}$ is the variation in intensity values and $\boldsymbol{\sigma}_{gr_3}$ is the variation in gradient angle[2] in a 3×3 window. Their product at every patch location in a 7×7 window is added. Intensity of a patch is normalized before calculations. The zoom value is directly proportional to the randomness measure. High intensity variations and low angle variation imply ramp like structure. High angle variation and low intensity variation imply noise. Higher value of both imply higher zoom factor. To identify ridge like structure as regular structure, gradient angle is computed in the range $[-\pi/2, \pi/2)$ instead of full 2π range. Proposed randomness measure fails to identify step edges because just after the steep, gradient angle could be anything in presence of noise. These edges are removed from consideration using canny edge detector. Proportional to the randomness measure zoom value is assigned as successive integer levels and Markov network is initialized. Fig. 4(b) shows some of the patch structures and the estimated zoom values.

[2] $dx = x_{t+1} - x_t$, $dy = y_{t+1} - y_t$, $\theta = \tan^{-1}(dy/dx)$

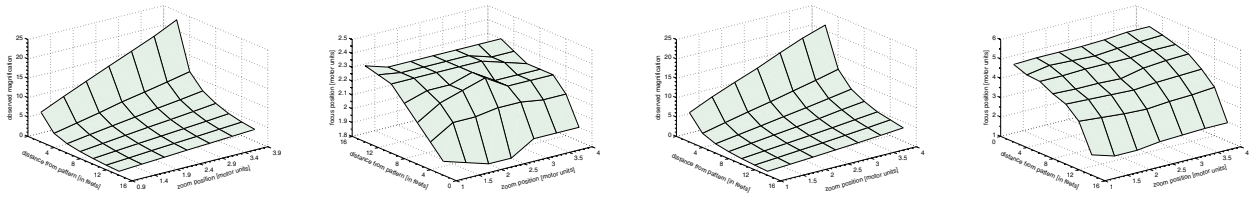
4. Calibration of Zoom Lenses

"Zoom lens model" defines the relationship between focus, zoom and aperture values. Scene magnification is controlled by moving two or more lenses along the axis and point of focus is selected by moving the whole lens assembly to and fro. The functional relationship between the various zoom lens parameters is obtained empirically rather than mathematically [23] because of high complexity of zoom lenses, unavailability of specifications of lenses and missing markings of zoom and focus motor position. We predict zooms upto 5X in experiments. We use two zoom lenses with focal length in the range 18-55mm and 28-105mm because of unavailability of a single high zoom lens. Virtually a $105/18 \approx 6X$ zoom lens is available. A high precision scale is affixed on focus and zoom motors. Images of checkerboard pattern are captured as a function of distance (in feets) and zoom position (in motor units). Homography matrix is computed among between the base image and other images of the pattern. Average of scale factor along two axis is the effective magnification. Fig. 5 shows the calibration graphs. Coupling table 1 is made which defines the relationship between two zoom lenses. It is the magnification achieved at minimum focal length by changing the zoom lenses.

Setting the Right Zoom: Let the first image is captured at (z_i, t_i), z_i and t_i denote the zoom and focus motor position respectively. Let m denote the required magnification factor and (z_f, t_f) denote the desired zoom lens configuration. If zoom motor position is fixed in the graph, then the focus position is monotonous as a function of distance from the pattern. This result follows from the fact that only one depth point of the scene remain in focus. Let $M(,)$ and $F(,)$ are the magnification profile(Fig. 5(a),5(c)) and focus profile(Fig. 5(b),5(d)) of the lenses. If M_k^{-1} and F_k^{-1} are the inverse of M and F at a constant k. The required zoom-lens parameters (z_f, t_f) are obtained as,

$$d = F_{z_i}^{-1}(t_i),$$
$$z_f = M_d^{-1}(mM(d, z_i)),$$
$$t_f = F(z_f, d).$$

Intermediate values are computed by fitting higher order polynomials as described in [23]. Coupling table is used to switch to other zoom-lens and the equations are similar.



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

Figure 5. Zoom lens calibration (a) and (c): magnification profile of two cameras as a function of zoom motor position and distance of the camera plane from the checkerboard (measured in feets); (b) and (d) corresponding focus position in motor units.

Distance (in feets)	2	4	6	8	10	12	14
Magnification	1.5617	1.5755	1.5695	1.5724	1.5770	1.5898	1.5832

Table 1. Coupling table : computed at the minimum focal length between two lenses as a function of distance.

Image	MSE (initialization)	MSE (MAP-MRF)
Book-shelf	0.3453	0.2671
Butter-fly	0.2921	0.2424
Bill-board	0.3672	0.2938
Book-text	0.3156	0.2398
Painting	0.2801	0.2250

Table 2. Evaluation results on synthetic data. Mean square error (MSE) is computed between the actual resolution-front value and computed using a) randomization measure, b) MAP-MRF.

5. Experiments and Results

To evaluate the performance of the proposed algorithm experiments are performed on a variety of real and simulated data-sets. As data is lost near boundary, several constraints are proposed to minimize the extent of zoom. Later in this section, several possible applications are also discussed. Around 54 images are selected which can be super-resolved further using simple SR algorithms. Randomization measure defined in section 3.6 is also used to check the suitability of training images. The size of the training patch is 8×8. It has 4×4 pixels from the underlying patch and other pixels from neighboring patches. Around 110,000 training patches are generated. Each training image is downsampled at various factors upto 8. 4-5 of these images are chosen and resolution-front values of them are computed. Patches are stored in the training database. Any learning based SR algorithm can be used during training phase (denoted as algorithm **A** in Fig. 3). [14, 8] are recent such algorithms. References therein provide further details on various similar algorithms. This algorithm is used in our experiments. Zoom is predicted upto $5X$ at intervals of 0.25. The desired zoom of the camera is calculated from the predicted resolution-front value. It is done by finding a largest rectangle (usually located at the center) for which the maximum resolution-front value is less than or equal to the size of the image divided by the size of the rectangle.

Performance on Synthetic Data: We take test images and downsample it. The resolution-front of each of them is computed as described in section 3.4 and also using our algorithm. The comparison with initialization and prediction in MAP-MRF framework is summarized in table 2.

Results on Real Data[3]**:** We first evaluate the performance on Snellen eye chart, which has various random alphabets printed at different font size. Fig. 6 summarizes the result. At various locations in Fig. 6(c) the resolution-front values are highly regularized. Whereas in Fig. 6(b) regions around the text are also marked for zoom. Prior information learned from the training data (e.g. regions above and below text should require no zoom) was useful. Fig. 6(f), 6(g) and 6(h) are images captured at increasing zoom. Various characters are clear at different zoom-levels.

Fig. 7 summarizes results on a slightly complex scene. Contextual constraints was very helpful in regularizing resolution-front values. In some of the cases, the final character size after zooming is slightly different. This is primarily because of different font types. Also predicting high zoom values from limited data could be slightly erroneous.

5.1. Constrained Zoom-in

To minimize the data loss near outer boundary of an image, several constraints are introduced. Scene is zoomed-in upto a level only if the constraints are met.

Visually Attentive Objects : To speed up many computer vision algorithms, certain regions are preferentially processed based on their visual attentiveness. This constraint is used to preferentially treat a region which is visually attentive. Publicly available 'Saliency Toolbox' which implements the algorithm by Walther and Koch [22] is used to locate such regions. Fig. 8 summarizes the results.

Penalty for not Zooming-in : The zoom is costly if a few blocks require very high zoom. A graph is constructed on zoom factor versus number of blocks requiring zoom factor greater than various zoom factors. Graph is normal-

[3] all images in this section should be enlarged to view them properly.

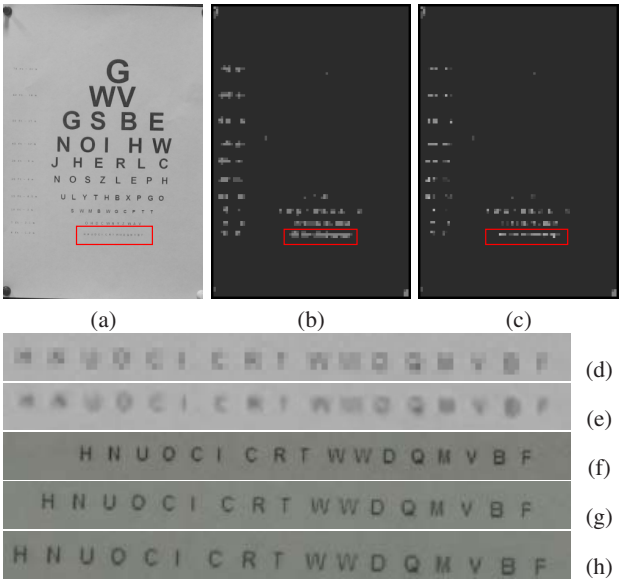


Figure 6. Experiments on Snellen chart (a) base image (b) zoom predicted using randomness measure with maximum zoom value 3 in the selected region (c) resolution-front predicted after optimizing equation 6 having values 3, 3.25, 3.5 and 4 in the selected region (d) selected region scaled by a factor 4 (e) super-resolved region; same patch after capturing images at zoom: (f) 3X (g) 3.5X (h) 4X.

ized and the first zoom factor where the value falls below a threshold is selected. It is also helpful to cope noise in resolution-front prediction. Fig. 9 summarizes the results.

Other Scenarios : Pre-determined objects can be segmented and such image regions is kept at higher priority. Separating man-made and natural structures [13] also provide useful constraints for zooming. Natural objects and scenes have fine details but they convey very little useful information. Whereas man-made object usually do not have intricate structures. Natural structures can also be replaced with any high quality texture while super-resolving.

5.2. Applications

Integration of the proposed technique with professional or consumer cameras can provide a simple way to capture high-quality images. The algorithm can also be used to predict required magnification factor for multi-frame SR algorithms. Robotics and surveillance systems require the interpretation of scenes which are usually unknown. It is impossible to scan scenes at maximum zoom value. Given that the most of the scene information in real world do not convey meaningful information or do not require very high zoom values. The scene can be captured optimally with minimum number of images at right zoom. For large scale image mosaicing (e.g. giga-pixel camera [12]) such algorithms can optimize the number of images captured. In automated cropping systems, regions which require very high zoom values can be removed after predicting the resolution front using our algorithm. As images captured at right zoom has almost all the information for further resolution enhancements, consequently the recognition accuracies of many systems will improve. For many real-time applications e.g. video surveillance, two camera systems can be used. One for capturing the whole scene and the other to capture only certain regions in detail.

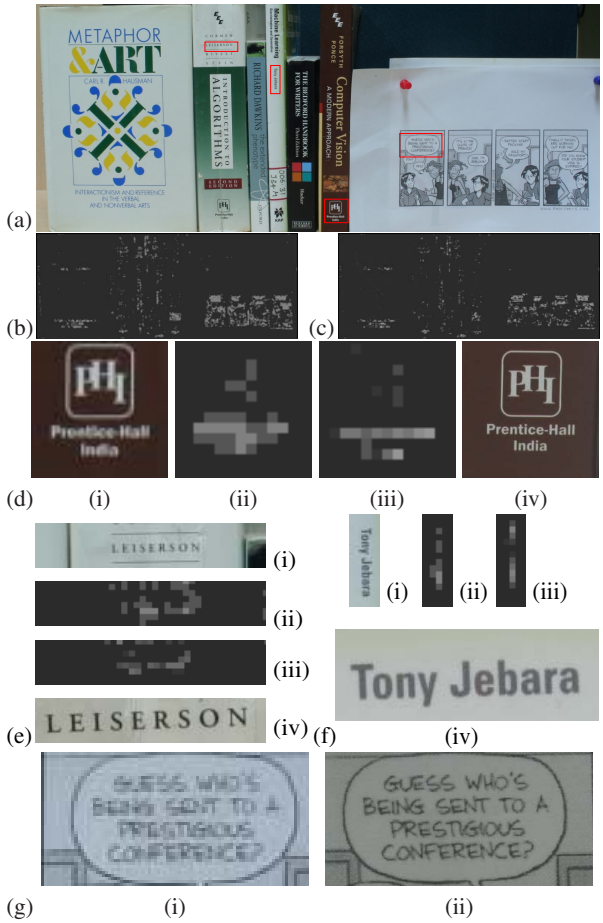


Figure 7. (a) base image (b) zoom predicted using randomness measure (c) resolution-front predicted after optimizing eq. 6; (d), (e) and (f): (i) selected regions from image (ii) initial resolution-front (iii) resolution-front after optimization (iv) regions shown at right zoom with values (d.iv) 3.5X (e.iv) 2.5X (f.iv) 2.5X (g.ii) 2.5X.

6. Discussions and Conclusions

In this paper, we have presented and addressed the problem of capturing the right amount of scene information from the perspective of SR. The final captured image can be mag-

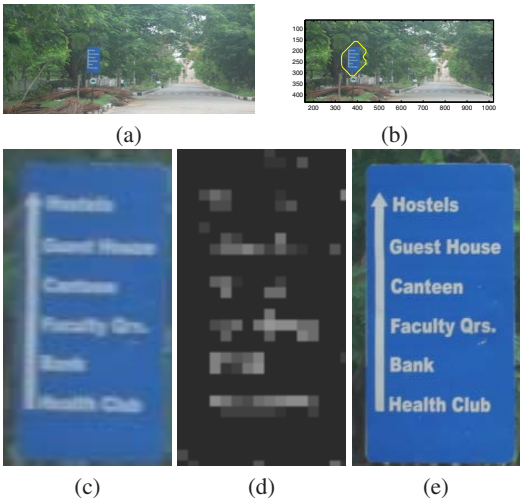


Figure 8. (a) base image (b) visually attentive region selected using saliency toolbox (c) selected LR region (d) R_f predicted (e) at right zoom (2.5X).

nified further using any learning based SR algorithm. The solution is proposed in a MAP-MRF framework. MRF allows modeling of contextual constraints. In Fig. 7(d.ii) and 7(f.ii) the initialized resolution-front values are not consistent and correct. With contextual constraints much of the regularization is brought in and resolution-front values are suppressed at unusual places. Places where underlying patch information is insufficient to predict high zoom values, context information played a significant role. In Fig. 8(e), the vertical line and the text have almost similar structure but the presence of context information is able to define right resolution-front values at various places. Selecting the right zoom value can as well be proposed as a high-level vision problem where a particular object is zoomed in at a pre-defined value. Proposing it as a low-level vision problem provides high degree of generalizability for a variety of scenes. Low computational speed is one of the key issues. But with additional constraints (section 5.1) significant speed up has been achieved. Camera shakes introduce blur in images and deteriorates the zoom prediction. But it can be controlled in autonomous environments. Future work is towards developing complete real-time systems for zoom prediction. We also plan to address the problem of locating useful structures in images. We envision that such a functionality would be introduced in consumer cameras.

References

[1] K. Aoyama. Auto-zoom camera. *US Patent No. 5604562*, 1997.

[2] O. Arandjelovic and R. Cipolla. A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution. In *ICCV*, 2007.

[3] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[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, September 2002. (To Appear).

[5] A. Belahmidi and F. Guichard. A partial differential equation approach to image zoom. *Image Processing, 2004. ICIP '04. 2004 International Conference on*, 1:649–652, Oct. 2004.

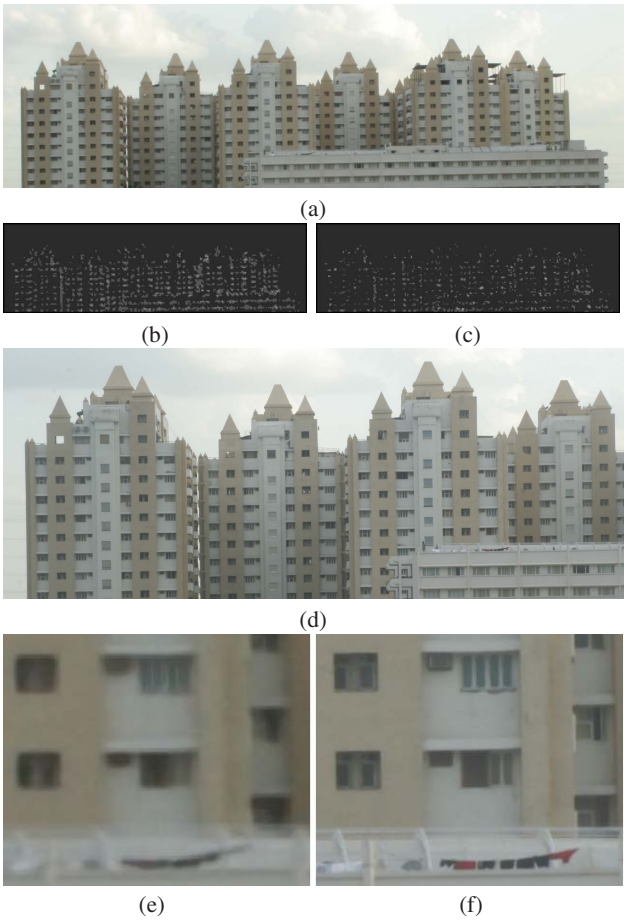


Figure 9. (a) base image (b) resolution-front initialization (c) resolution-front predicted (d) image captured at 2.5X zoom. Highest resolution-front value was 4.25X; (e) super-resolved image of (a) by 5X; (f) super-resolved image of (d) by 2X. Many structures are clear in (d).

[6] J. E. Bollman, R. L. Rao, D. L. Venable, and R. Eschbach. Automatic image cropping. *US Patent No. 5978519*, 1999.

[7] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *Image Processing, IEEE Transactions on*, 6(12):1646–1658, 1997.

[8] R. Fattal. Image upsampling via imposed edge statistics. *ACM Trans. Graph.*, 26(3):95, 2007.

[9] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000.

[10] T. Hashimoto, M. Ikemura, K. Kimura, Y. Hata, K. Hayashi, H. Ootsuka, and M. Nakanishi. Camera having an auto zoom function. *US Patent No. 5291233*, 1994.

[11] R. Jin, Y. Qi, and A. Hauptmann. A probabilistic model for camera zoom detection. In *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 3*, page 30859, Washington, DC, USA, 2002. IEEE Computer Society.

[12] J. Kopf, M. Uyttendaele, O. Deussen, and M. F. Cohen. Capturing and viewing gigapixel images. In *SIGGRAPH '07: ACM SIGGRAPH 2007 papers*, page 93, New York, NY, USA, 2007. ACM.

[13] S. Kumar and M. Hebert. Man-made structure detection in natural images using a causal multiscale random field. In *in proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 119–126, 2003.

[14] X. Li and M. Orchard. New edge-directed interpolation. *Image Processing, IEEE Transactions on*, 10(10):1521–1527, Oct 2001.

[15] Z. Lin, J. He, X. Tang, and C. Tang. Limits of learning-based superresolution algorithms. In *ICCV*, pages 1–8, 2007.

[16] Z. Lin and H.-Y. Shum. Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1):83–97, 2004.

[17] J. Luo. Automatically producing an image of a portion of a photographic image. *US Patent No. 6654507*, 2003.

[18] J. Luo and R. T. Gray. Method for automatically creating cropped and zoomed versions of photographic images. *US Patent No. 6654506*, 2003.

[19] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab. *Signals & systems (2nd ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[20] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *Signal Processing Magazine, IEEE*, 20(3):21–36, 2003.

[21] B. Tordoff and D. Murray. Resolution vs. tracking error: zoom as a gain controller. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Madison, Wisconsin*. IEEE Computer Society Press, June 2003.

[22] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Netw.*, 19(9):1395–1407, 2006.

[23] R. Willson. Modeling and calibration of automated zoom lenses. In *Proceedings of the SPIE No. 2350: Videometrics III*, pages 170 – 186, October 1994.