Attention-Based Super Resolution from Videos

Dileep Vaka, P. J. Narayanan and C. V. Jawahar
Center for Visual Information Technology, IIIT Hyderabad.
{dileep@research., pjn@, jawahar@} iiit.ac.in

Abstract

A video from a moving camera produces different number of observations of different scene areas. We can construct an attention map *of the scene by bringing the frames to a common reference and counting the number of frames that observed each scene point. Different representations can be constructed from this. The base of the attention map gives the scene mosaic. Super-resolved images of parts of the scene can be obtained using a subset of observations or video frames. We can combine mosaicing with super-resolution by using all observations, but the magnification factor will vary across the scene based on the attention received. The height of the attention map indicates the amount of super-resolution for that scene point. We modify the traditional super-resolution framework to generate a varying resolution image for panning cameras in this paper. The varying resolution image uses all useful data available in a video. We introduce the concept of attention-based super-resolution and give the modified framework for it. We also show its applicability on a few indoor and outdoor videos.*

1. Introduction

Videos in general contain multiple observations of the scene. The number of observations for different parts of regions vary based on the attention received. For any video, a mosaic can be built by combining information from different frames. The repeated observations can be used to construct a high resolution image using super resolution. Traditional super resolution expects the entire scene to be visible in each low resolution observation. The magnification factor depends on the number of such observations. Only a part of the scene in the video can be super resolved using this strategy. In contrast, we would like to use the maximum information from the video. Combining mosaicing and super resolution. As the mosaic is built, each region is super resolved based on the attention received by it. For example, a panning video has more observations at the center of the mosaic than at the edges. We must magnify the center part of the mosaic by a higher factor compared to the edges, to utilize the complete available information.

Several image mosaicing methods have been proposed earlier [13, 14, 17]. They align or register images using correspondences between them. Image registration is a well studied problem in [3, 13]. The overlapping portions of the aligned images may be averaged or blended together for creating a mosaicing. Super resolution is a well explored problem and it has been achieved using different ways [11, 9, 6, 8]. Limits of the super resolution have also been explored from a theoretical point of view [1].

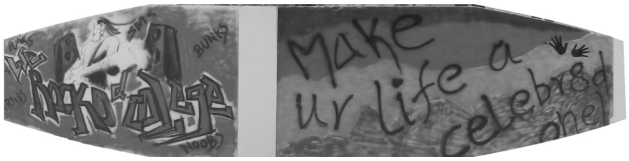

Figure 1. An attention-based super resolution mosaic. The resolution varies continuously from the left and right edges to the center.

Teodosio and Bender present the Salient Stills representation that has a wide field of view and high resolution for regions of intrest [15]. They construct such a representation from a video that starts as a long shot and zooms in to object of intrest. Our work aims to automatically deduce the object of intrest from from the attention given in the videos.

There have been attempts to super resolve mosaics in the past [5, 4, 18]. The previous works on super resolved mosaic have focused on alignment problems and different ways of solving for a high resolution image. The focus was not on the optimal magnification factor for super resolution. In [18] the super resolved mosaics are obtained using generalized strips. The overlapping frames are aligned to strips and each strip is super resolved independently. This approach can super resolve each strip to a different magnification factor and join them together, but the objects may look discontinuous at the joining of the strips.

This paper focuses on building a super resolved mosaic

978-0-7695-3476-3/08 $25.00 © 2008 IEEE
DOI 10.1109/ICVGIP.2008.101

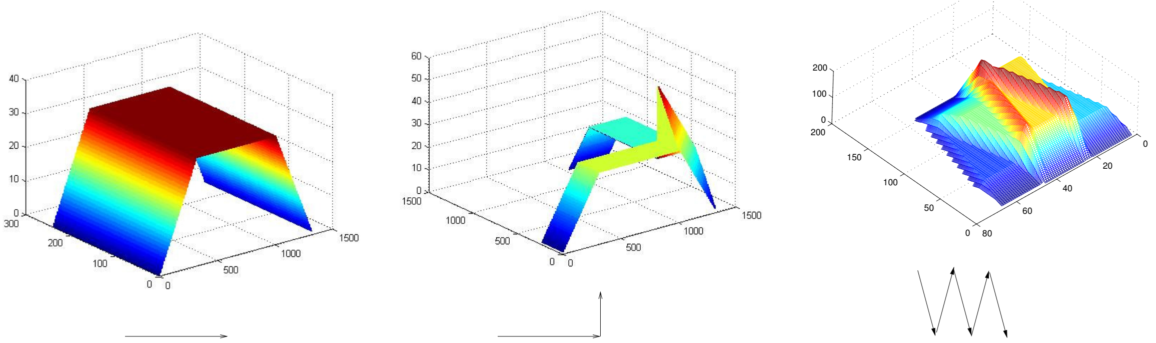

Figure 2. Attention Map for (left) horizontal panning, (middle) horizontal panning followed by tilting and (right) zig-zag camera motion. Camera motion for these attention maps are shown below the attention maps.

that magnifies each particular region by a factor proportional to the number of samples available in that region as shown in Figure 1. We first build an attention map which shows the number of samples for each pixel in the mosaic. Using the attention map we build a super resolved image with a varying magnification factor. The traditional super resolution process is modified to accommodate varying magnification factors without image discontinuities. We also define attention maps for different kinds of videos and study their relation with super resolution and image mosaicing.

2. Attention Map of a Video

A video from a moving camera produces different number of observations of different scene areas. We can analyze the attention received by each scene region with the help of an *attention map*. An attention map gives the number of observations received by each scene region in the video. It can be built by bringing all the frames into a common reference and counting the number of video frames that observe each scene region. The video can be used to construct different representations based on attention map. The attention map is a 3D plot with scene regions (or their projections to the camera) as its base. The height gives the number of observations of each scene region.

The left most image in Figure 2 shows the attention map of a camera panning horizontally. The number of observations increase from the left most part of the visible scene, stabilizes to a maximum value, and falls to the right of the scene. The slope depends on the camera speed. The plateau in the middle represents the region that received the maximum attention. For a constant panning velocity the plateau will start at the right most end (or the width) of the first frame and end at the left most column of the last frame.

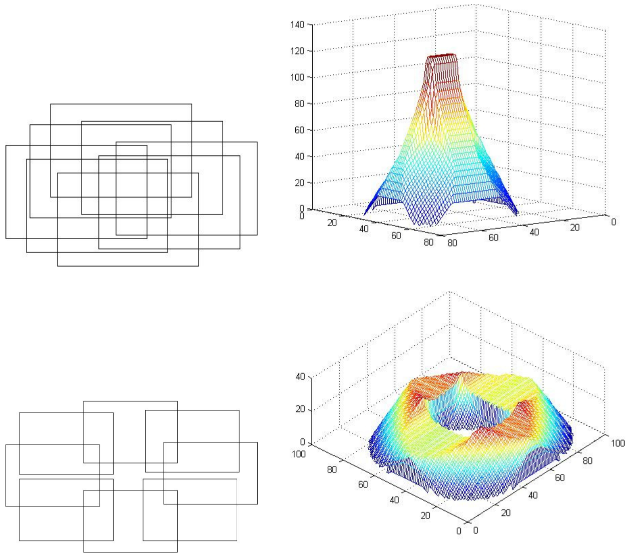

Figure 3. Attention map for case of (top) high overlap at the center and (bottom) no overlap at the center.

The attention map of a horizontal pan followed by vertical tilt will have a sudden rise as shown in center image of Figure 2. Right most image of the Figure 2 shows the attention map for a zigzag camera motion. Figure 3 shows the attention map for a rotating camera motion. Top image in Figure 3 has the central region visible in all frames. Its attention map has a clear peak at the center and in the bottom image of Figure 3 the center is not visible in any frame and the attention map has a hole in the center.

The attention map can be understood as follows. The shape of the mosaic that can be built from the video is the projection of the attention map onto its base. The observations at each scene point can be combined to give the appearance in the mosaic of the projection of that region. The combining is performed by blending or by taking the mean of the overlapping regions.

When the camera is moving continuously, the volume under the attention map of a video is a measure of the total attention received by all the scene regions. Since the magnification factor for super resolution for a region is proportional to the number of observations received (or its square root in two dimension), the changing heights indicate the varying magnification factor of the attention-based super resolved mosaic that can be built. We can directly map the volume under the attention map to the area of the final attention-based super resolution mosaic built from the video.

Previous researchers have studied the issue of overlap of different parts of the image for building mosaics of larger field of view [5, 12]. The objective has been to provide a hole-free coverage of the scene while mosaicing. The attention-map differs from them as it is a quantitative representation of the amount of attention received by scene regions. This enables its use for simultaneously mosaicing the scene and super-resolving it by a factor decided by the input data.

3. Varying Super Resolved Mosaics

Videos with panning will contain multiple samples of the scene across time. These frames can be registered to a common reference frame and a mosaic that expands the field of view can be built or a super resolved image can be constructed by combining multiple observations. Super resolution is done for a fixed magnification factor for the whole scene. When different parts of the scene have different number of observations, super resolution has to use the common minimum number of observations for the scene region of interest. This tends to ignore the additional samples at parts of the scene that are observed many times. We intend to use all samples to construct a super resolved image. This combines mosaicing with super resolution, with a varying magnification factor over the mosaic depending on the number of samples available.

Our goal is to apply an magnification factor to each region of the scene based on the attention map. In Section 3.1 we will look at overview of super resolution and in Section 3.2 we will establish the correspondence between the mosaic and the varying super resolution image, and then we construct varying super resolution image.

3.1. Super Resolution: Overview

Super resolution [10, 2] is a method to construct an image with higher spatial resolution than the original one. First step of the super resolution construction is to formulate a model that relates a high resolution (HR) image to the low resolution (LR) images. Let us denote the measured LR images by Y_i and the image formed by super resolving these images by X. These images are converted into column vectors by lexicographical ordering, so that matrix operations can be done over them. Each LR image is assumed to be formed after the super resolved image undergoes geometric warp, blur and downsampling. We can write the super resolution equation as

$$\overrightarrow{Yi} = D_i B_i T_i(\overrightarrow{X}) + \eta_i, \tag{1}$$

where T_i is the geometric warp operation between X and Y_i, B_k is the blurring matrix, η_i is the additive noise and D_i is the downsampling matrix. From the above equation HR image can be obtained using different approaches [4, 7].

Super resolution reconstruction techniques can be divided into frequency domain [16] and spatial domain. To compute the observation model there exists several techniques in literature like Maximum Likelihood (ML), Maximum a Posterior (MAP) estimation method, and Projection onto Convex Sets (POCS). An iterative way of solving for super resolution is to minimize the error between low resolution image and the simulated low resolution image. Iterative solution using

$$X^{j+1} = X^j + H^{BP}(Y - Y^j), \tag{2}$$

where Y^j is the simulated LR image and H^{BP} is the back projection operator, will converge to the super resolved image X.

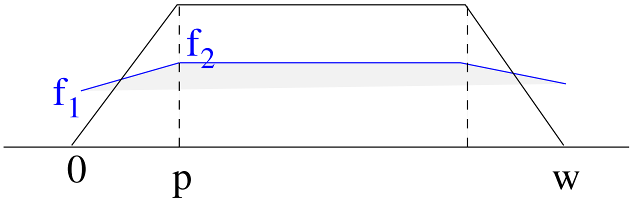

Figure 4. Black line indicates the attention map across a row of the mosaic and the blue line indicates the magnification factor.

3.2. Varying Super Resolution of Videos

For a video, each region can be super resolved to a factor that depends on its attention map. Magnification of each region is proportional to the height of the region in the attention map. We reformulate the super resolution process to directly create a varying resolution image. The shape of the output image will depend on the shape of the attention map. In this work, we assume that the resolution varies linearly with the height of the attention map. Thus, the volume of the attention map will define the shape of the varying resolution mosaic. The change in resolution has to be effected without creating discontinuities in the output image. We take a simple case of horizontal panning and look at building the variable resolution mosaic from the panned video.

The pixels in the mosaic will get super resolved to a factor proportional to the attention map value for that pixel.In Figure 4, blue line indicates the varying magnification factor and the black line indicates the attention map. Say a_1 and a_2 be the minimum and maximum number of samples from the attention and let f_1, f_2 be their respective super resolving factors. If a is the number of observations available at any arbitrary point in the mosaic then it should be super resolved by a factor of $\simeq \frac{a-a_2}{a_1-a_2}(f_2 - f_1) + f_1$. The factor f_1 will be 1 if a_1=1 as there is a point in the mosaic which has only one sample. And f_2 will depend on the highest resolution possible. If the different regions of the mosaic are magnified with varying magnification factors then the output varying image will look like Figure 1.

The mapping between the mosaic and the varying super resolution image has to be established as the primary step. The mapping between them is not a direct scaling as the magnification factor varies across the image. Let p be the width of the frames in the video and w be the width of the mosaic. The magnification factor corresponding to the attention map will increase linearly from 1 to f in the region 1 to p in the horizontal direction. So a point in the mosaic whose x-coordinate x is less than p will get super resolved by a factor of $\frac{x-1}{p-1}(f-1)+1$. The x coordinate of this point in varying super resolution resolution image will lie at $\sum_{i=1}^{i=x} \frac{i-1}{p-1}(f-1)+1$, as it is equal to the sum of magnification factor of the points lying to the left of it. Width of the super resolved image will thus be the sum of the magnification factors

$$2\sum_{i=1}^{i=p}(\frac{(i-1)(f-1)}{p-1}+1)+(w-2p)f, \tag{3}$$

which equals to $p(f+1)+f(w-2p)$. Where as width of the super resolved mosaic with a constant magnification is wf.

Varying super resolved images can be obtained by incorporating the varying magnification factor into the super resolution framework, we assume that a varying super resolution image underwent warping, blurring and downsampling to get various LR observations. Normal super resolution formulation is still applicable because the variation in magnification factor will be small among the neighboring pixels in the mosaic. So the local neighborhood of any pixel in the mosaic will be super resolved by similar factors. Since the super resolution construction depends only on the neighboring pixels the normal super resolution construction is applicable. We modify the normal super resolution in the following way. A varying super resolution image on warping, blurring and downsampling gives us a transformed mosaic, which does not completely overlap with the observed frames. So after downsampling, we multiply it with a selection matrix S_i (which contains 0's and 1's) to give us the observed i^{th} low resolution image.

$$\overrightarrow{Yi} = S_i D_i B_i T_i \overrightarrow{(X)} + \eta_i \tag{4}$$

Down sampling matrices can be calculated using the correspondence established previously. We can use the normal super resolution solving techniques to solve this equation. Maximum likelihood approach to our problem would be

$$L(\overrightarrow{X}) = \frac{1}{2} \parallel (\overrightarrow{Y} - A\overrightarrow{X}) \parallel \tag{5}$$

where A is the combined linear operation of S, D, B and T. Differentiating L w.r.t X

$$A^T(\overrightarrow{Y} - A\overrightarrow{X}) = 0 \tag{6}$$

$$\sum_{i=1}^{n} T_i^T B_i^T D_i^T (D_i B_i T_i X - Y_i) = 0 \tag{7}$$

Simplest way to solve this is by using steepest decent algorithm. The steepest descent algorithm suggests the following iterative equation for the solution of above equation.

$$\widehat{X_{j+1}} = \widehat{X_j} + \lambda \sum_{i=1}^{n} T_i^T B_i^T D_i^T (D_i B_i T_i X - Y_i) = 0 \tag{8}$$

X_0 is the initial estimate of the super resolution. The upsampled mosaic be taken as the initial estimate.

4. Results and Discussion

We tested the attention-based super resolution on different panning videos. Middle image in the Figure 5 shows the output of a varying super resolution created from the video of 640 × 480 which had 60 frames in it. We calculated homographies between every two consecutive frames using Harris corners. The images were then aligned with respect to the first frame. Using the aligned images, an attention map was built. Maximum and minimum heights of the surface in attention map were 26 and 1 respectively. Based on attention map the shape of the varying super resolution image was decided. The mosaic and the varying resolution images were built. Top and bottom images in the Figure 5 show the mosaic and the super resolved images obtained by magnifying the whole mosaic with a constant magnification factor of 2.

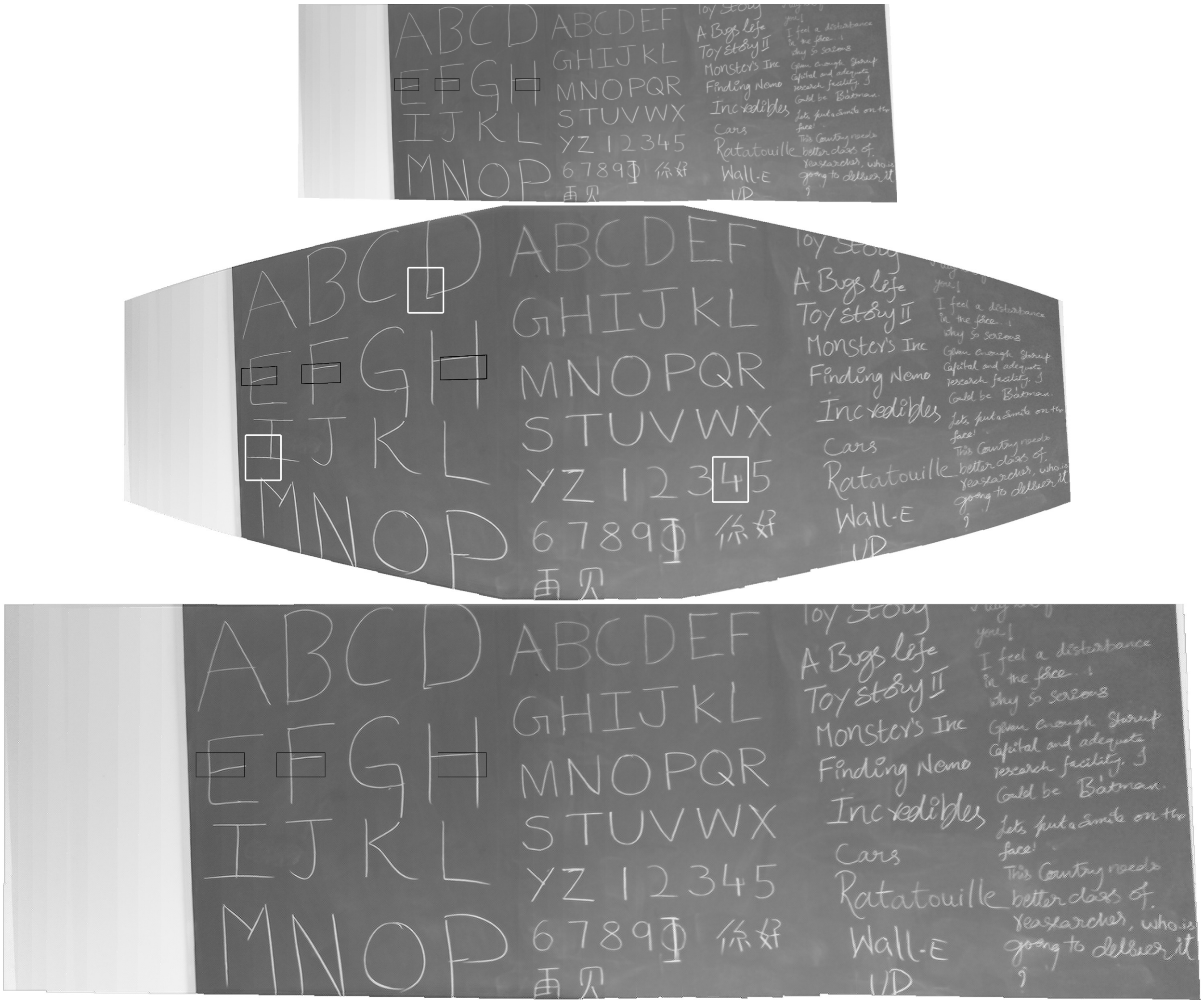

Figure 5. (top) Mosaic of the video sequence (1505×491). (middle)Attention based super resolved image output with a highest magnification factor of 2 (2370×982). (bottom) Mosaic super resolved by a constant factor (3010×982).

In Figure 6, we compared regions of highlighted rectangles from the middle and bottom images of Figure (5). The resolution of the varying resolution image at the edges is less when compared to super resolved mosaic with a constant magnification factor, as the magnification factor in the attention based super resolution image is close to 1 at the beginning of the mosaic. Mosaic super resolved with constant magnification is similar to the scaled image of the mosaic at the edges. As we move right the quality of the constant super resolved image improves, where as the quality of the varying super resolution at the edges is as good as at the center.

Figure 7 shows the comparison between the varying super resolution image and the mosaic for the black bordered rectangles shown in Figure 5. We can observe that the regions in the mosaic that got magnified by different factors in the varying resolution image are at same quality. Figure 8, shows the the regions shown in white of the varying resolution image . We can notice that the quality of all the three regions is almost same and the text size of the center image is higher than the other two, but the text size were almost of same size in mosaic. This shows us that center region got super resolved by a higher factor than the other regions.

Top image in Figure 9 shows the output of a varying super resolution created from a outdoor video of buildings. Resolution of the video was 640×480 and it had a minimum overlap of 1 and maximum overlap of 30. Middle and bottom images in the Figure 9, show the output of varying super resolution created from a video of resolution 320×240, which had a maximum overlap of 32 and 37 respectively.

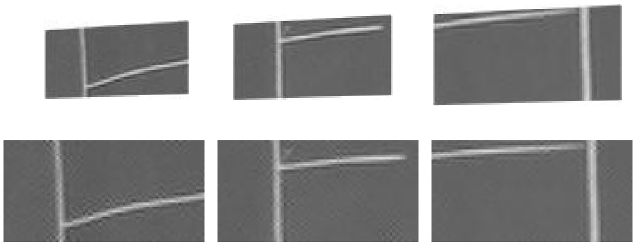

Figure 6. Comparison between the varying resolution image (top) and the mosaic super resolved by a factor of 2 (bottom) for the three black bordered windows show in Figure 5.

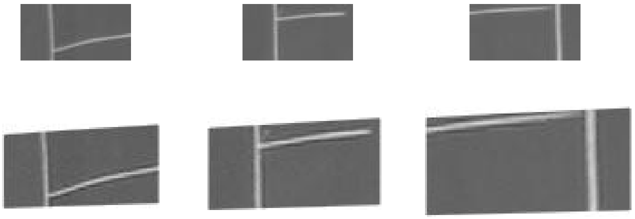

Figure 7. Comparison between the mosaic (top) and the varying resolution image (bottom) for the three black bordered windows shown in Figure 5.

5. Conclusions and Future Work

In this paper, we proposed the concept of the attention map, which quantifies the amount of observation received by different scene regions of a video. This information was used in simultaneously super-resolving and mosaicing the video. The magnification factor depends on the attention received, which varies across the mosaic. We computed varying super-resolution mosaics for different panning videos.

We intend to extend this idea to videos with independently moving objects, which may receive different amounts of attention. This calls for different representations and super-resolution techniques.

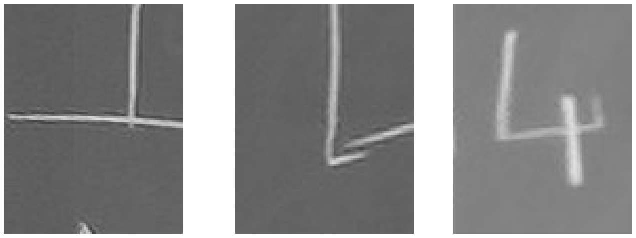

Figure 8. Equal sized windows from three parts of the varying resolution image show in white in Figure 5.

References

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. In *CVPR*, pages 372–379, 2000.

[2] S. Borman and R. Stevenson. Spatial resolution enhancement of low-resolution image sequences - A comprehensive review with directions for future research, Nov. 12 1998.

[3] Brown. A survey of image registration techniques. *CSURV: Computing Surveys*, 24, 1992.

[4] D. Capel. Image mosaicing and super-resolution. In *Ph.D.*, 2001.

[5] D. Capel and A. Zisserman. Automated mosaicing with super-resolution zoom. In *CVPR*, pages 885–891, 1998.

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

[7] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar. Advances and challenges in super-resolution. *International Journal of Imaging Systems and Technology*, 14(2):47–57, 2004.

[8] M. Irani and S. Peleg. Super resolution from image sequences. In *ICPR*, pages II: 115–120, 1990.

[9] M. Irani and S. Peleg. Improving resolution by image registration. *Graphical Models and Image Processing*, 53:231–239, 1991.

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

[11] D. Rajan and S. Chaudhuri. Generalized interpolation and its application in super-resolution imaging. *Image and Vision Computing*, 19(13):957–969, Nov. 2001.

[12] H. S. Sawhney, S. C. Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. In *ECCV (2)*, pages 103–119, 1998.

[13] H.-Y. Shum and R. Szeliski. Systems and experiment paper: Construction of panoramic image mosaics with global and local alignment. *International Journal of Computer Vision*, 36(2):101–130, 2000.

[14] R. Szeliski and H.-Y. Shum. Creating full view panoramic image mosaics and environment maps. pages 251–258, 1997.

[15] L. Teodosio and W. Bender. Salient video stills: Content and context preserved. In *ACM Multimedia*, pages 39–46, 1993.

[16] R. Tsai and T. Huang. Multiframe image restoration. In *Advances in Computer Vision and Image Processing*, pages 317–339, 1984.

[17] I. Zoghlami, O. Faugeras, and R. Deriche. Using geometric corners to build a 2d mosaic from a set of image. In *Proceedings of CVPR*, page 420, 1997.

[18] A. Zomet and S. Peleg. Efficient super-resolution and applications to mosaics. In *ICPR*, pages 1579–1583, 2000.

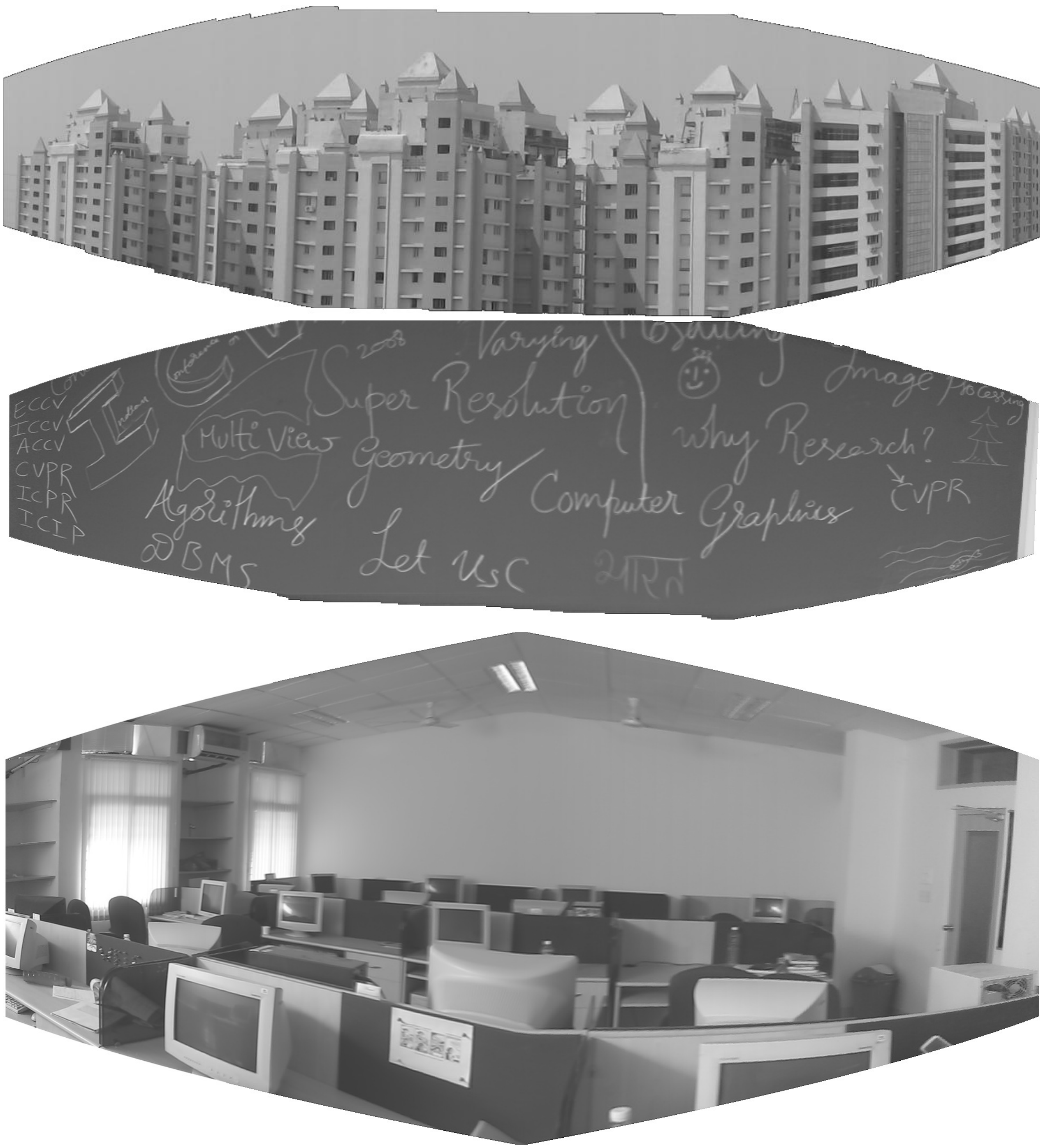

Figure 9. Varying resolution images created from a horizontal panning of (top) buildings, (middle) some text and (bottom) a indoor scene.