High-Resolution Multiscale Panoramic Mosaics from Pan-Tilt-Zoom Cameras

Sudipta N. Sinha Marc Pollefeys Seon Joo Kim
Department of Computer Science, University of North Carolina at Chapel Hill.

Abstract

We present an automatic software based approach for building extremely high resolution panoramic mosaics from images captured by an off-the-shelf pan tilt zoom camera. Using numerous zoomed-in images from such a camera, existing mosaicing algorithms could theoritically build gigapixel images. However the large number of images that must be processed makes such approaches impractical. Our stratified approach for high resolution panoramic imagery, first constructs a coarse panorama of the scene and then adds in detail from a zooming camera, in a top-down multi-resolution fashion. Our approach uses both feature based and direct intensity based image alignment methods. Both the geometric calibration (intrinsic parameters and radial distortion) as well as the photometric calibration and alignment is done automatically. Our fully calibrated panoramas are represented as multi-resolution pyramids of cubemaps. We align hundreds of images captured within a 1-12X zoom range and show results from two datasets captured from cameras placed in an uncontrolled outdoor scene.

1. Introduction

Omnidirectional cameras use special sensors to simultaneously image a scene with a large field of view (FOV) often from a single viewpoint. Such cameras typically capture low-resolution images and have a limited range of scale. High-resolution panoramic cameras require special hardware and can be extremely expensive. Static scenes however do not require simultaneous imaging. Hence multiple images, (each with a small FOV) captured over time can be aligned and composited into a complete panorama using image mosaicing algorithms [1, 2, 8, 9]. Such methods require an overlap between adjacent images, and are impractical for building high resolution full-view mosaics as they cannot efficiently handle a large number of images.

We propose to use affordable off-the-shelf pan-tilt-zoom (PTZ) cameras to construct high resolution panoramic mosaics. These cameras by virtue of their large zoom range can view a scene at a range of scale much larger than an omnidirectional camera. At its finest scale, it can capture high-resolution imagery whereas a large range of pan and tilt gives it a large virtual FOV. Hence the PTZ camera combines the best of both worlds at an affordable cost.

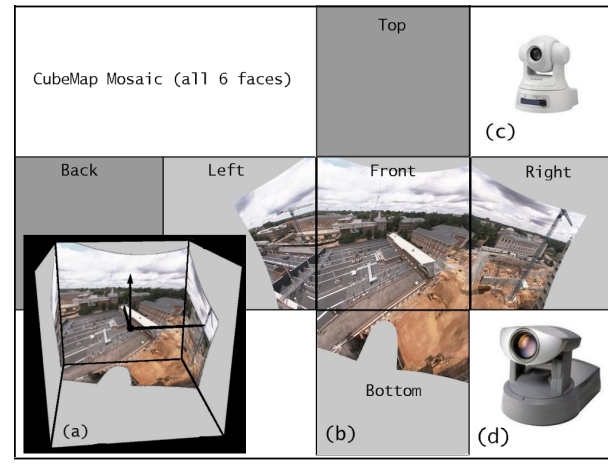


Figure 1: Cube Map Mosaic of 84 images from a PTZ Camera. (a) mapped on a cube.(b) 6 faces unfolded on a plane. PTZ Cameras - (c) Sony SNC-RZ30 (d) Canon VB-C10.

In our coarse to fine approach, first a coarser version of the mosaic is computed by stitching overlapping images captured from a rotating camera at a small fixed value of zoom. During this step, the intrinsics, radial distortion parameters and the photometric calibration of the PTZ camera are also determined. Next the camera repeatedly sweeps its FOV with increasing zoom, thus acquiring images of the same scene with more and more detail. These images are independently aligned with an existing mosaic to produce a version with higher resolution. Adjacent images do not require much overlap, and hence for a particular scene, a minimum number of images are processed. The camera zoom is doubled for each acquisition phase and higher resolution mosaics are computed iteratively to form a multi-resolution pyramid of cubemap mosaics. The faces of a cubemap (see Fig. 1(a,b) represents an omnidirectional image from a camera whose projection center is at the center of the cube. This representation is suitable as images from a purely rotating and zooming camera are related by a $2D$ homography [4].

We use two types of PTZ Cameras, the Canon VB-C10 and Sony SNC-RZ30 (see Fig. 1). They have large pan and tilt range, 16X-25X maximum optical zoom, 300K-400K pixel CCDs with horizontal FOV ranging from 47^o to 2^o-3^o

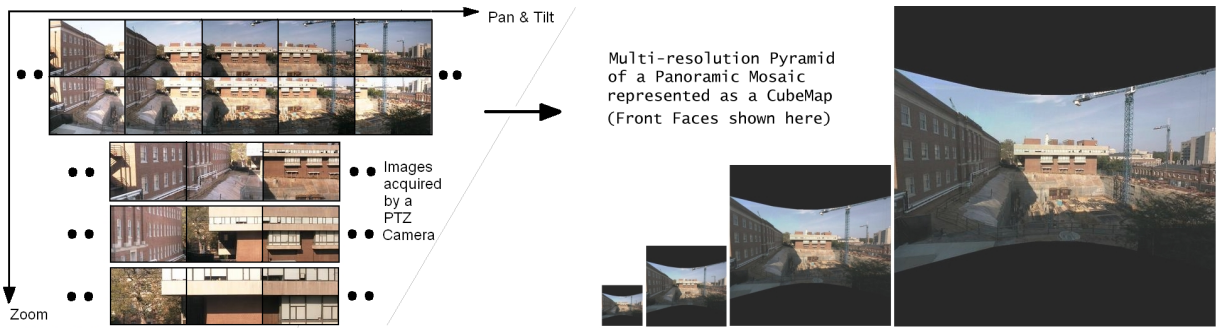


Figure 2: Multi-resolution pyramid of a cubemap mosaic from images captured by a zooming and rotating PTZ camera.

when fully zoomed in. The Canon and Sony cameras can capture upto 1.12 and 3.09 gigapixels at maximum optical zoom respectively. However the PTZ controls of these cameras are not repeatable and precise calibration can only be obtained from the images themselves [10]. Our method relies on robust sub-pixel image alignment [5, 7]. We use robust feature-based methods [3, 4] initially but adopt direct intensity-based methods [5] for accuracy once the images are roughly registered and photometrically aligned. The dense sampling of rays present in high-resolution mosaics is key to high fidelity 3D reconstruction of wide-area environments, image-based rendering and activity detection at multiple levels of detail in surveillance systems. Section 2 discusses the relevant theory while our multi-resolution approach is described in Sections 3. We present experimental results in Section 4 and conclude with scope for future work in Section 5.

2. Theory and Background Work

2.1 Camera Model

We chose a simple pin-hole camera model where the camera's center of rotation is fixed and coincides with the center of projection while it is rotating and zooming. Such an assumption is valid, when the PTZ camera is used outdoors or in large environments where the shift of the camera center is small compared to its distance to the observed scene. Our experiments have shown that the Canon VB-C10 and Sony SNC-RZ30 surveillance cameras follow this model with reasonable accuracy. In the pin-hole model (see Fig. 3) for the perspective camera, a point $\mathbf{X}$, in $3D$ projective space P^3 projects to a point $\mathbf{x}$, on the $2D$ projective plane P^2 (the image plane). This can be represented by a mapping $f : P^3 \rightarrow P^2$ such that $\mathbf{x} = \mathbf{PX}$, $\mathbf{P}$ being the 3×4 rank-3 camera projection matrix (see Eq. 1).

$$\mathbf{P} = \mathbf{K}[\mathbf{R} - \mathbf{Rt}] \;\; \mathbf{K} = \begin{pmatrix} \alpha f(z) & s & p_x(z) \\ 0 & f(z) & p_y(z) \\ 0 & 0 & 1 \end{pmatrix} \quad (1)$$

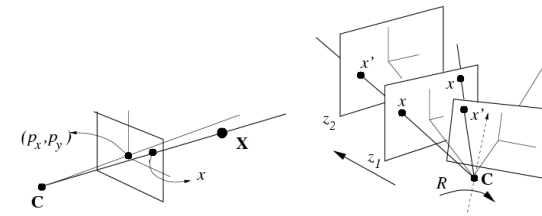


Figure 3: Left : The pin-hole camera model. Right : Camera undergoing pan-tilt rotation and zoom.

where $\mathbf{K}$ represents the camera intrinsics while $\mathbf{R}$ and $\mathbf{t}$ represents the camera orientation and position in the world coordinate system. The matrix $\mathbf{K}$ can be expressed in terms of α, s, f, p_x and p_y (Eq. 1), where α and s are the camera's x:y pixel aspect ratio and skew (we assume zero skew, hence s=0); f its focal length in pixels, (p_x,p_y) its principal point and z its current zoom.

Most cameras deviate from a real pin-hole model due to radial distortion which becomes more prominent for shorter focal lengths. The $3D$ point $\mathbf{X}$ which projects to $\mathbf{x} = (\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \mathbf{1})$ under the pin-hole model actually gets imaged at (x_d,y_d) due to radial distortion as shown in Eq. 2.

$$\begin{pmatrix} x_d \\ y_d \end{pmatrix} = \mathbf{L}(\tilde{\mathbf{r}}) \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} \quad (2)$$

$\tilde{r} = \sqrt{\tilde{x}^2 + \tilde{y}^2}$ is the radial distance of $\mathbf{x}$ from the center of distortion (x_c,y_c) (assumed to coincide with the principal point) and $\mathbf{L}(\tilde{\mathbf{r}})$ is a distortion factor determined by $\tilde{r}$. The function $\mathbf{L}(\mathbf{r})$ is represented as $\mathbf{L}(\mathbf{r}) = \mathbf{1} + \kappa_\mathbf{1}\mathbf{r}^\mathbf{2} + \kappa_\mathbf{2}\mathbf{r}^\mathbf{4}$ and (κ_1, κ_2) is the parametric model for radial distortion. For a PTZ Camera, the focal length f , the principal point (p_x,p_y) and coefficients of radial distortion (κ_1 and κ_2) are functions of the zoom. A method for computing the parameters over the camera's full zoom range is described in [10].

2.2 Rotating and Zooming Cameras

Here we consider the case of a rotating and zooming camera. Let x and x' be the images of $\mathbf{X}$ taken at two different instants by a camera that is either zooming or rotating or both. These points, x and x' are related to X as $x = \mathbf{K}[\mathbf{R}\ \mathbf{t}]\mathbf{X}$ and $x' = \mathbf{K}'[\mathbf{R}'\ \mathbf{t}]\mathbf{X}$ where $t = 0$. Hence, $x' = \mathbf{K}'\mathbf{R}'\mathbf{R}^{-1}\mathbf{K}^{-1}\mathbf{x}$. In our model, the intrinsics remain the same for pure rotation at constant zoom, and hence this equation reduces to $x' = \mathbf{K}\mathbf{R_{rel}}\mathbf{K}^{-1}\mathbf{x}$ where $\mathbf{R_{rel}} = \mathbf{R}'\mathbf{R}^{-1}$ represents the relative camera rotation about its projection center between the two views and $\mathbf{K}$ is the camera intrinsic matrix for that particular zoom level. Similarly for a zooming camera with fixed center of projection, $x' = \mathbf{K}'\mathbf{K}^{-1}\mathbf{x}$. These homographies are represented by $\mathbf{H_{rot}}$ and $\mathbf{H_{zoom}}$ (see Eq. 3).

$$\mathbf{H_{rot}} = \mathbf{K}\mathbf{R}\mathbf{K}^{-1} \qquad \mathbf{H_{zoom}} = \mathbf{K}'\mathbf{K}^{-1} \tag{3}$$

3. The Multi-resolution Approach

Conventional mosaic algorithm [2, 8, 9] would be infeasible for stitching hundreds of images all captured at high-zoom to build extremely high resolution full-view mosaics. For instance, assuming 50% overlap between adjacent images, the SNC-RZ30 must capture 21,600 images at 25X zoom (full FOV of 3.16π steradians) while the VB-C10 needs 7800 images at 16X zoom (full FOV of 2.55π steradians). By adopting a coarse to fine multi-resolution scheme, where images captured at a particular zoom are aligned with a mosaic built from images at half the present zoom, approximately half of the above image count would be needed at full zoom. The multi-resolution framework itself does require additonal images to be captured at intermediate zooms. However by using a top-down pruning scheme, we reduce the number of images captured by avoiding high zoom in areas where detail is absent (see Section 3.3).

Figure 4(a) gives an overview of our approach. Phase I, dealing with building the base cubemap C_0 (for the lowest zoom) and geometric calibration of the camera, is described in [10]. Section 3.1 describes the extension to include photometric calibration. This allows a consistent blending of the base cubemap. Phase II outlined in Fig. 4(b) involves building a cubemap, C_z of size $2N$ x $2N$ pixels from images captured at zoom level z using the cubemap C_{z-1} of size N x N computed previously from images at roughly half the zoom. The recorded pan p_j^z, tilt t_j^z associated with every captured image I_j^z is used to generate an image from the calibrated cubemap C_{z-1} at half the resolution. For a perfectly repeatable camera, these two images denoted by A and B in Fig. 4(b) should be perfectly aligned. The SNC-RZ30 and VB-C10 however require additional alignment because of the inherent non-repeatability of the PTZ controls. First a feature based method [4] (Chap.3, page 108)

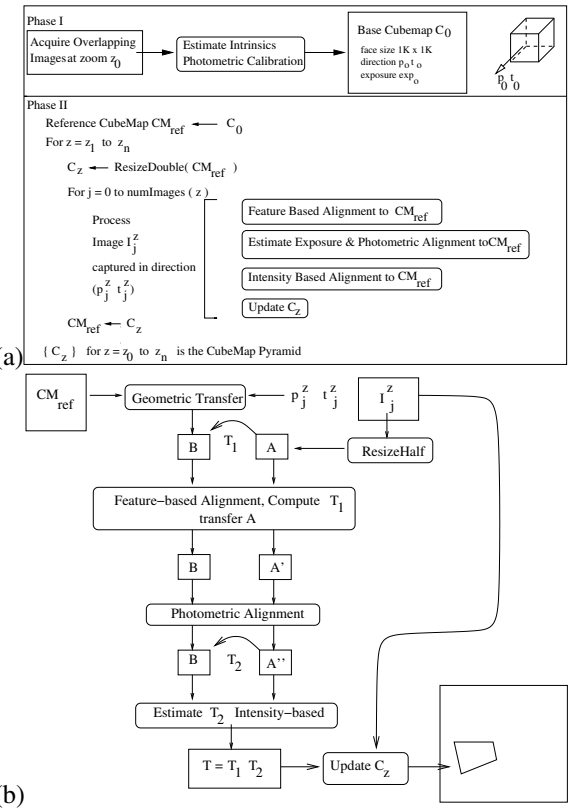


Figure 4: (a) Overview of our method. (b) Phase II : Coarse to Fine Cubemap pyramid construction.

is employed which works in the presence of outliers as well as intensity changes in the two images. Once the images are roughly aligned, the exposure of this new image, A' is estimated (see Section 3.1). Now accurate and robust alignment is performed by an intensity based method described in Section 3.2. Once the cubemap at zoom level z is built, its becomes the base cubemap for the next level. Every level of the cubemap pyramid is initialized from the previous level by bilinear interpolation.

3.1. Robust Radiometric Calibration

The intrinsic calibration of a PTZ camera of [10] is extended here to include photometric calibration. The camera senses a high range of radiance (brightness intensity) in the scene while acquiring images in auto-exposure mode. Hence the captured images have different exposures and must be transformed to a common exposure before they can be blended into a single mosaic. The camera's response function is robustly estimated from the overlapping images captured at its lowest zoom in Phase I and the exposures of all the images are computed using the method described in [6]. The pixel correspondences required by this method are obtained from an accurate sub-pixel registration of all the

images, [10] based on Bundle Adjustment [11]. Once the camera's response function is known, the exposure of every subsequent zoomed-in image captured in Phase II can be estimated using the same method after registering it to a mosaic of known exposure. The results of blending the stitched images after photometric alignment is shown in Fig. 5(a,b).

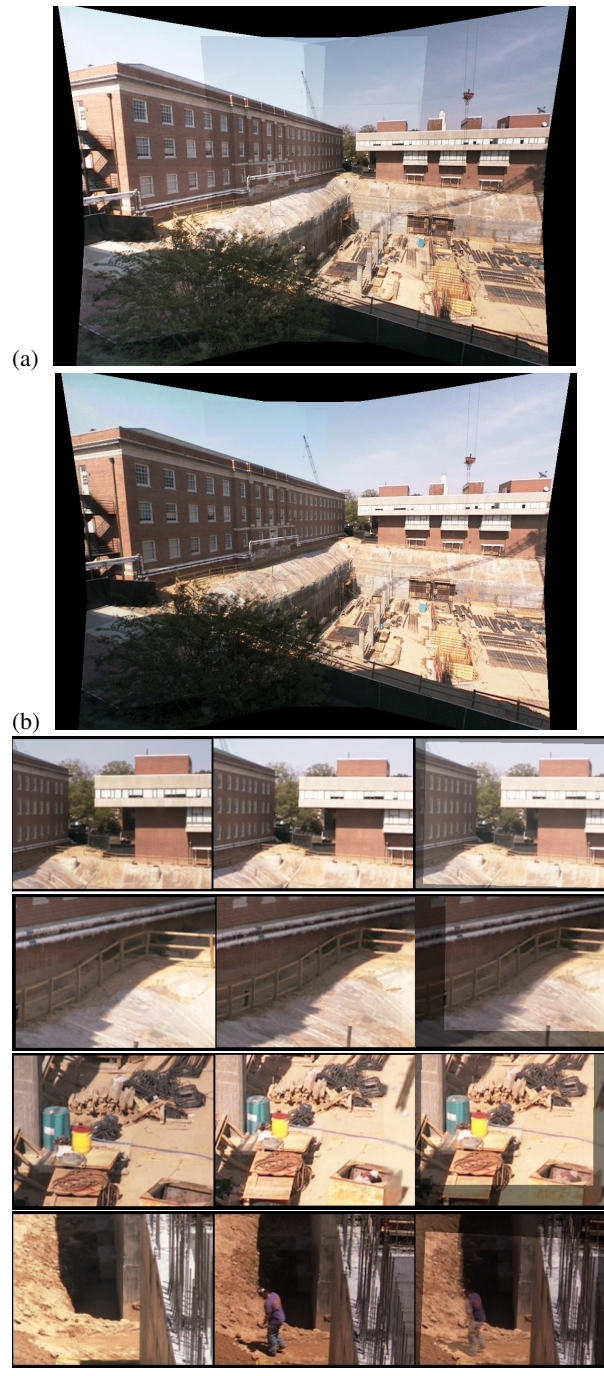
3.2. Image Alignment

Accurate sub-pixel image alignment is key to building accurate mosaics. Phase I and the initial alignment step for every image in Phase II of our method uses an implementation of the RANSAC-based [3] robust estimation algorithm described in [4] (Chap.3, page 108). In the presence of sufficient reliable corners, an affine homography is estimated. However when reliable corners are absent, the feature-based method falls back on computing homographies with fewer degrees of freedom ie. a similarity transformation or a translation. When the image contains substantial non-rigid motion, for eg. fairly zoomed in images of moving branches and leaves, this step typically fails and a direct intensity based registration is attempted.

During Phase II, a direct intensity based method is used to improve the registration between the current image and the reference cubemap. A Coarse-to-fine Lukas Kanade optic flow estimation [7] is computed followed by a RANSAC step to fit an affine model to the observed optic flow in the presence of outliers. Hence even when images contain moving objects like vehicles, trees, branches, leaves etc. the static portions of the scene are registered accurately since they satisfy an affine flow whereas the moving objects are treated as outliers. The presence of moving objects, moving shadows could be removed as described in [5]. Small motion at a coarser scale gets accentuated when the camera zooms in and hence alignment fails beyond a certain zoom level. Fig. 5(c) shows some examples where the RANSAC-based affine flow computation produces a good registration of the static parts of the scene even in the presence of shadows that move with time and moving objects and people.

3.3. Image Acquisition

The computational infeasibility of directly constructing a high resolution mosaic was described in Section 3. Building the mosaic pyramid in a coarse to fine fashion requires multiple acquistion passes, which captures the scene at a range of scales. This requires us to inspect images at a coarser scale (low zoom) to decide which parts of the scene contain detail. Often large portions of the scene contain textureless regions, for eg. the sky, walls, roads. We avoid zooming into such homogeneous regions and reduce the number of image acquired considerably. In order to complete acquisi-



(a) (b) (c)

Figure 5: Front face of a base cubemap rendered (a) without photometric alignment and (b) with photometric alignment. (c) Image Alignment: The 3 columns show the captured frame, the corresponding image generated from the cubemap and the aligned image pair (the first one overlaid on the second) respectively. In the $2nd$ and $4th$ images, in spite of moving shadows (images were taken far apart in time), the static parts of the scene are accurately aligned.

.

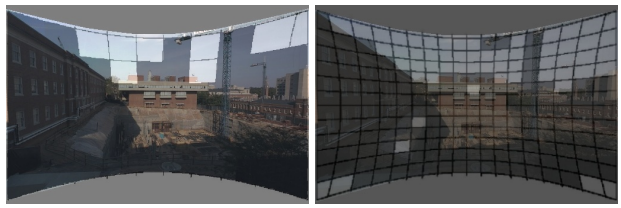


Figure 6: Shaded regions on the base mosaic (1X zoom), indicating where images were captured at zoom levels 4X and 8X respectively. Regions like the sky were skipped when the camera zoomed in.

tion quickly, we do not wait to first build the calibrated base cubemap before subsequent passes at higher zoom. Instead an approximate calibration is used to backproject pixels into rays and effectively decide on the basis of texture analysis, whether the image at a specific PTZ value should be captured or skipped. An image block, where the eigen values of its second moment matrix are large, is mapped to a ray using the corresponding pan and tilt values, which is inserted into a kd-tree [1]. While scanning the scene in the next pass, a range query within this kd-tree returns a ray-count within the camera's FOV. Viewing directions corresponding to a low count contain mostly textureless regions in the scene. These images are skipped at the current and subsequent zoom levels. Our approach will miss texture present at finer scales which are filtered at coarser scales. However this allows us to directly acquire a compressed version of a very high resolution image instead of acquiring a raw image and then compressing it using lossy techniques. The result of pruning at two higher zoom levels is shown in Fig. 6.

4. Implementation and Results

We built two cubemap pyramids, one each from images captured by a Sony SNC-RZ30 and a Canon VB-C10 camera placed outdoors looking at a construction site (see Fig. 7). The 1024 x 1024 pixel (face size) base cubemaps were built by stitching 15 and 9 overlapping images respectively. In each case the multi-resolution pyramids had five levels upto a resolution of 16K x 16K pixels. The camera captured 15-20 images ie. 5-6.5 Mpixels at 3X zoom. About 70-95 images were captured at 6X zoom, which produced 21.5-29 Mpixels. Finally 300-350 images were captured at 12X zoom, out of which 200-250 were successfully aligned and hence contributed 62-77 Mpixels. These unique pixels in addition to the pixels interpolated from lower levels in the pyramid made up the faces of all the cubemaps. Scene1 and Scene2 (Fig. 7) were processed in 1-1.5 hrs on a 1.5 GHz Notebook Computer with 512 MB RAM. Each of the original images were 640 x 480 pixels (1:10 compressed jpg). In our implementation for the multi-scale mosaic pyramid construction, we used a tile-based representation for the large cubemap faces and processed them out-of-core using a image tile cache with FIFO block replacement strategy. This implementation is scalable and can potentially create gigapixel images for full-view panoramas by processing upto a few thousand images to build six full cubemap faces at $16K$ x $16K$ pixel resolution.

5. Conclusions and Future Work

An algorithm to construct full-view, high resolution cubemap mosaics using a rotating and zooming PTZ camera is presented. Our coarse to fine approach based on robust image alignment builds a multi-resolution pyramid representation of the cubemap. Currently our cameras capture individual images, hence acquisition is slow. A faster video based capture will be explored in future in order to build 1+ gigapixel images. This will allow motion segmentation of moving objects which can be removed automatically resulting in more accurate background mosaics.

Acknowledgements

This work was supported by the NSF Career award IIS 0237533.

References

[1] M. Brown and D. Lowe. Recognizing panoramas. In *ICCV03*, pages 1218–1225, 2003.

[2] D. Capel and A. Zisserman. Automated mosaicing with super-resolution zoom. In *CVPR98*, pages 885–891, 1998.

[3] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In *SRI-TN*, 1980.

[4] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. In *Cambridge*, 2000.

[5] M. Irani and P. Anandan. About direct methods. In *Proceedings of the International Workshop on Vision Algorithms*, pages 267–277. Springer-Verlag, 2000.

[6] S. Kim and M. Pollefeys. Radiometric alignment of image sequences. In *CVPR04*, 2004.

[7] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 1981 Darpa Image Understanding Workshop*, pages 121–130, 1981.

[8] H. Sawhney and R. Kumar. True multi-image alignment and its application to mosaicing and lens distortion correction. In *CVPR97*, pages 450–456, 1997.

[9] H. Shum and R. Szeliski. Systems and experiment paper: Construction of panoramic image mosaics with global and local alignment. *IJCV*, 36(2):101–130, February 2000.

[10] S. Sinha and M. Pollefeys. Towards calibrating a pan-tilt-zoom camera network. In *OMNIVIS04*, 2004.

[11] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, 2000.

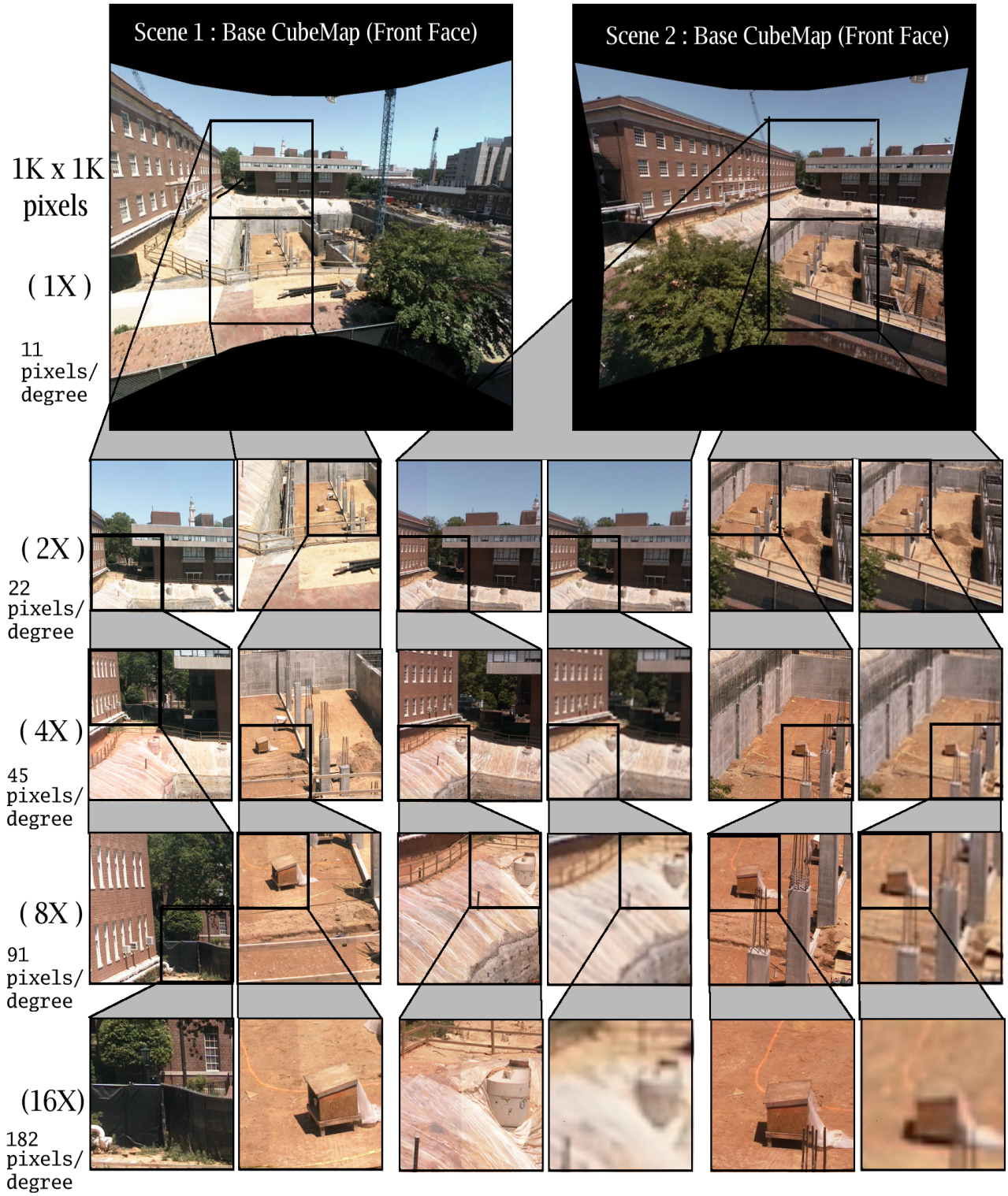


Figure 7: Results: Two Cubemap pyramids with 5 levels were built, where each face had $1K$, $2K$, $4K$, $8K$ and $16K$ sq. pixels. Certain zoomed-in sections (512 x 512 actual image pixels) are shown above. Column 1 and 2 shows the Level of Detail for two parts of Scene 1. Column 3 and 5 shows two parts of Scene 2 at different level of detail. Compare the resolution with Column 4 and 6 showing the same view enlarged from the $1K$ x $1K$ base cubemap.