Photorealistic 3D Modeling of Architecturally Complex Environments

Ioannis Stamos,
Department of Computer Science, Hunter College, City University of New York, NY 10021,
istamos@hunter.cuny.edu

Abstract

This paper presents a system that is suitable for reconstructing large and complex urban environments. This becomes possible by the development of novel algorithms for 3–D model acquisition from the combination of range and image sensing. The input is a sequence of unregistered range scans of the scene and a sequence of unregistered 2-D photographs of the same scene. The output is a true texture-mapped geometric model of the scene. The vital parts of the system (segmentation, range registration,solid modeling, and texture mapping) are presented. Segmentation algorithms simplify the dense data-sets and provide stable features of interest that can be used for registration purposes. Range registration and solid modeling provides geometrically correct 3-D models. Finally, automated range to image registration algorithms can increase the flexibility of the system by decoupling the slow geometry recovery process from the image acquisition process.

1 Introduction

The recovery and representation of 3–D geometric and photometric information of the real world is one of the most challenging and well studied problems in Computer Vision and Robotics research. There is a clear need for highly realistic geometric models of the world for applications related to Virtual Reality, Telepresence, Digital Cinematography, Digital Archeology, Journalism, and Urban Planning. Recently, there has been a large interest in reconstructing models of outdoor urban environments [13]. The areas of interest include geometric and photorealistic reconstruction of individual buildings or large urban areas using a variety of acquisition methods and interpretation techniques, such as ground-base laser sensing, air-borne laser sensing, ground and air-borne image sensing. The ultimate goal is the reconstruction of detailed models of urban sites (digital cities). The creation of digital cities drives other areas of research as well: visualization of very large data sets, creation of model data-bases for GIS (Geographical Information Systems) and combination of reconstructed areas with existing digital maps.

The problem we attack can be described as follows: Given a set of dense 3-D range scans of a complex real scene from different viewpoints and a set of 2-D photographs of the scene, a) create the 3-D solid model that describes the geometry of the scene, b) recover the positions of the 2-D cameras with respect to the extracted geometric model and c) photorealistically render it by texture-mapping the associated photographs on the model. The integrated system we developed for the production of photo-realistic geometric models of large and complex scenes is described in figure 1.

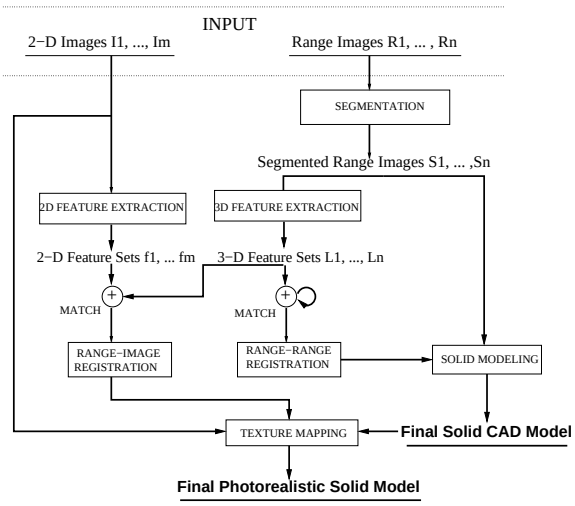

Figure 1: System for building geometric and photometric correct solid models.

This paper provides an overview of our photorealistic 3D modeling approach [18, 20, 19, 1] and introduces new results from the 3D reconstruction of St. Piérre Cathedral in Beauvais, France (section 6). Related work is presented in section 2 . Section 3 provides and overview of our segmentation and solid modeling approaches, section 4 describes the range registration algorithm, whereas section 5 presents the range-image registration method.

2 Related Work

There are two major approaches in the photorealistic reconstruction of urban 3-D scenes: purely geometric (extraction of dense geometry via range sensing or sparse and irregular geometry via stereo techniques) and image-based rendering methods (extrapolating geometry in the rendering phase via resampling the captured light field of the scene). Representative systems whose goal is the photorealistic reconstruction of real scenes

by the utilization of 2-D imagery only are [7, 17, 2]. In those cases the necessary human interaction and the a-priori geometric constraints imposed by the human operator lead to lack of scalability with respect to the number of processed images of the scene and to the computation of simplified geometric descriptions of the scene. Teller's approach [21] addresses the limitations of the previously described methods by acquiring and processing a large amount of pose–annotated high-resolution spherical imagery of the scene.

Systems that extract dense and regular geometry *must* rely on accurate range measurements. Representative approaches include the VIT group [22, 3], the Digital Michelangelo project [14], the Pietá Project [4], Fitzgibbon et. al. [11], Zhao [23] and Sequiera [16]. Finally Zisserman's group in Oxford [12] works towards the fully automatic construction of Graphical Models of scenes when the input is a sequence of closely spaced 2–D images (video sequence). This work shows how far purely image–based methods have gone but also points out the following inherent limitations: sparse depth estimates which depend on the texture and geometric structure of the scene, and crude surface approximations in areas that do not support 3–D measurements.

3 Segmentation & Modeling

The individual range-images which the range-sensor provides are the result of dense sampling of visible surfaces in large urban scenes. Using a Cyrax laser scanner [6], we get 1K by 1K range samples ($\sim$ 1 million range samples) with a spatial resolution of a few centimeters. Our first task is to segment the dense range scans and extract major surface patches. The segmentation is sequential-labeling type algorithm of the 3-D points into 8-connected regions, with a metric of similarity (co-planarity and co-normality) between neighboring points. This algorithm has complexity $O(N)$ where N is the total number of range points (in our experiments $N \sim 10^6$). At a second level of abstraction the 3–D range data-set is represented as a set of 3–D curves. Those curves are the result of intersection of neighboring bounded 3–D surfaces which have been extracted by the range segmentation module. We implemented the extraction of 3–D lines as a result of planar surface intersections. Those 3–D features are used for the registration between 3–D data-sets and between 3–D data-sets and 2–D images [19]. Finally, volumetric solid models are constructed from registered segmented scans. The registration between individual scans is based on manual match between extracted linear featured between the scans. Our modeler is based upon earlier work by Reed and Allen [15]. The innovative principle of this approach is the representation of each individual range image with a solid volume. Figure 2 presents results from segmentation and modeling of real buildings.

4 Range-Range Registration

To create a complete description of a scene we need to acquire and register multiple range images. The registration (computation of the rotation matrix R and translation vector $\mathbf{T}$) between the coordinate systems of the n_{th} (C_n) and first (C_1) range image is possible via a matched set of 3–D features between the images. We have decided to use the infinite 3–D lines which are extracted using our segmentation algorithms (section 3) as our features of interest. A manual match between a small number of those features provides enough constraints that can lead in the computation of the rotation R and translation $\mathbf{T}$.

In detail the algorithm works as follows. The infinite 3-D lines that are automatically extracted from the dataset[1] can be represented by the pairs of the form $(\mathbf{n}, \mathbf{p})$, where $\mathbf{n}$ is the unit vector which corresponds to the direction of the line and $\mathbf{p}$ is a 3-D position which represents a point on the line. Note that this representation is not unique. There are two valid line directions $\mathbf{n}$ and $-\mathbf{n}$ and an infinite number of points $\mathbf{p}$ that lie on the line. We choose $\mathbf{p}$ to be one of the extracted endpoints of the line.

A solution for the rotation and translation is possible when at least two line matches are given. The rotation matrix R can be computed according to the closed form solution described in [9], page 523.

Lets assume that the lines

$$(\mathbf{n_i}, \mathbf{p_i}), i = 1 \ldots N$$

extracted automatically from one view do match up with the automatically extracted lines

$$(\mathbf{n_i}', \mathbf{p_i}'), i = 1 \ldots N$$

of the second view.

The rotation component of the transformation between the two view can be computed using the orientations $\mathbf{n_i}$ and $\mathbf{n_i}'$ of the matched 3-D lines. This is done via the minimization of the error function

$$Err(N) = \sum_{i=1}^{N} ||\mathbf{n_i}' - R\mathbf{n_i}||^2$$

where R is the unknown rotational matrix. The minimization of the above function has a closed-form solution when the rotation is expressed as a quaternion. The minimum number of correspondences for the computation of the rotation is two ($N = 2$). More lines can though be used in order to increase the robustness of the method. Note that in the above formulation we

[1] Our modules extract 3-D lines of finite extent. However, the extracted positions of the endpoints are not used for registration purposes because of the uncertainty in their determination.

assume that we have a correspondence between the directed vectors $\mathbf{n_i}$ and $\mathbf{n_i}'$. Otherwise the minimization would be formulated as

$$Err(N) = \sum_{i=1}^{N} ||\epsilon_i' \mathbf{n_i}' - \epsilon_i R \mathbf{n_i}||^2$$

where $\epsilon_i, \epsilon_i = \pm 1$. The latter formulation results in 4 possible solutions for the rotation matrix. However, knowledge of the matching directions reduces the number of solutions to one.

Solving for the translation vector $\mathbf{T}$ between the two views is an easy task as long the rotational matrix has been computed. Let us select two arbitrary points on the $i_t h$ line $< \mathbf{n_i}, \mathbf{p_i} >$ of the first view. Those points can be expressed as $\mathbf{a_1^i} = \mathbf{p_i} + t_1 \mathbf{n_i}$ and $\mathbf{a_2^i} = \mathbf{p_i} + t_2 \mathbf{n_i}$ where $t_j, j = 1, 2$ are two arbitrary real constants. Those two points have corresponding points which lie on the $i_t h$ line $< \mathbf{n_i}', \mathbf{p_i}' >$ of the second view. If we call those points $\mathbf{a_1^{i'}}$ and $\mathbf{a_2^{i'}}$ we can similarly express them as $\mathbf{a_1^{i'}} = \mathbf{p_i'} + t_1' \mathbf{n_i'}$ and $\mathbf{a_2^{i'}} = \mathbf{p_i'} + t_2' \mathbf{n_i'}$, where $t_j', j = 1, 2$ are two real constants which depend on the arbitrary selection of $t_j, j = 1, 2$. That means that the correspondence between the $i_t h$ lines of the first and second view provide us the following constraints:

$$\mathbf{a_1^{i'}} = R\mathbf{a_1^i} + \mathbf{T} \quad (1)$$

$$\mathbf{a_2^{i'}} = R\mathbf{a_2^i} + \mathbf{T} \quad (2)$$

or

$$\mathbf{p_i'} + t_1' \mathbf{n_i'} = R(\mathbf{p_i} + t_1 \mathbf{n_i}) + \mathbf{T} \quad (3)$$

$$\mathbf{p_i'} + t_2' \mathbf{n_i'} = R(\mathbf{p_i} + t_1 \mathbf{n_i}) + \mathbf{T} \quad (4)$$

When the rotation matrix R is known the above system of 6 equations is linear in the 7 unknowns (3 for the translation vector $\mathbf{T}$ and 4 for the real constants t_j and $t_j', j = 1, 2$). With two line matches the number of equations becomes 12 ($= 2 \times 6$) and the number of unknowns 11 ($= 2 \times 4 + 3$). That means that a minimum of two matched infinite 3-D lines provide enough constraints for the computation of the translation (when the rotation is known) through the solution of an over-constrained system of linear equations.

Results of the registration algorithm on three data sets are presented in figure 3.

5 Range-Image Registration

We provide a solution to the automated pose determination of a camera with respect to a range sensor without placing artificial objects in the scene and without a static arrangement of the range-camera system [20]. This is done by solving the problem of automatically **matching** 3–D & 2–D features from the range and image data sets. Our approach involves the utilization of parallelism and orthogonality constraints that naturally exist in urban environments in order to extract 3–D rectangular structures from the range data and 2–D rectangular structures from the 2–D images. Similar features are extracted from the 3-D range scans. By utilizing the RANSAC [10] framework a match between the 3-D and 2-D feature space is found. This match is used in order to solve for the position of the camera with respect to the 3-D model. The resulted texture-map is shown in figure 2f. We are moving in the direction of extending those methods in less constrained environments.

6 The Beauvais Project

We have started testing our algorithms on range data gathered from the St. Piérre Cathedral in Beauvais France (the Cathedral is in the UNESCO list of endangered world monuments). This project is sponsored by the Media Center of Art History and Archeology of Columbia University [8]. The author collaborates with the Robotics Laboratory of Columbia University that is responsible for the reconstruction of the final 3-D model. The recovery of the 3-D model of the building will greatly help in its structural analysis. A photorealistic model is also very useful for teaching the architecture of the Cathedral. A data-set of more than one hundred interior and exterior range scans and hundreds of photographs of the building was gathered in the summer of 2001.

Figure 5 shows some of our current results. The top rows contains two photographs of the exterior of the building. The second row presents four registered range scans of the same portion of the building (distinct range scans are represented with different colors), and a detail of that registration. The registration is done by manually matching automatically extracted linear features between the range scans. The linear features are the output of the range segmentation routines (section 3). the third row. Finally, in the last image of the figure, a texture-map using two photographs is shown. Each pixel in the resulting image gets color from one of the two images. Note, that as shown in figure 4 an intelligent decision must be made regarding the surface color at e. In the result shown at figure 5b, we are using the color of one of the two cameras without performing any color blending.

7 Conclusions

This paper presents a systematic approach to the problem of photo-realistic 3–D model acquisition from the combination of range and image sensing. A review of our segmentation, modeling, and registration algorithms is presented. We provide results utilizing data gathered from complex urban structures. A very important range-to-image fusion problem that still needs to be addressed is the blending of color images captured from overlapping viewpoints on the 3D model. The texture-map shown in figure 5b is binary; each texture-mapped

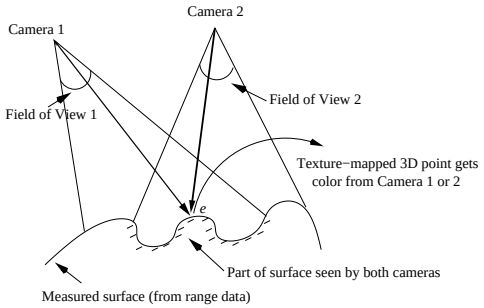

Figure 4: Texture mapping using more than one image. More than one camera views the same surface element e. The texture-mapping algorithm should decide which camera to use for texture-mapping e.

pixel gets its color from one of the two images. The next step is to intelligently blend the sequence of images that cover the whole scene. The main question here is which sets of images to use when a viewer looks at the scene from a particular viewpoint. In the recent work of [5], a unified framework in rendering a 3D scene using a large number of 2D images is presented. This framework is a generalization of view-dependent texture mapping [7] and light-field/lumigraph rendering approaches. A second issue that we still need to address is the complete automation of the range to range registration process.

Acknowledgments

We would like to thank Prof. Peter K. Allen of Columbia University and the Media Center of Art History and Archeology of Columbia University for using the Beauvais data-set. Finally, we would like to thank Dr. Andrew Miller for his help in the range-data acquisition on Beauvais.

References

[1] P. K. Allen, I. Stamos, A. Gueorguiev, E. Gold, and P. Blaer. AVENUE: Automated site modeling in urban environments. In *3rd Int. Conference on Digital Imaging and Modeling*, 2001.

[2] S. Becker. *Vision–assisted modeling from model–based video representations.* PhD thesis, Massachusetts Institute of Technology, Feb. 1997.

[3] J.-A. Beraldin, L. Cournoyer, et al. Object model creation from multiple range images: Acquisition, calibration, model building and verification. In *Intern. Conf. on Recent Advances in 3–D Dig. Imaging and Modeling*, pages 326–333, Ottawa, Canada, May 1997.

[4] F. Bernardini and H. Rushmeier. The 3D model acquisition pipeline. In *Eurographics 2000 State of the Art Report (STAR)*, 2000.

[5] C. Buehler, M. Bosse, L. McMillan, S. Gotler, and M. Cohen. Unstructured lumigraph rendering. *Computer Graphics (Proc. SIGGRAPH '01)*, pages 425–432, 2001.

[6] Cyrax technologies, 2000. http://www.cyra.com.

[7] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecure from photographs: A hybrid geometry-based and image-based approach. In *SIGGRAPH*, 1996.

[8] E. Eakin. Cybersleuths take on the mystery of the collapsing colossus. New York Times, October 27 2001.

[9] O. Faugeras. *Three–Dimensional Computer Vision.* The MIT Press, 1996.

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Graphics and Image Processing*, 24(6):381–395, June 1981.

[11] A. Fitzgibbon, D. Eggert, and R. Fisher. High-level cad model acquisition from range images. *Computer-Aided Design*, 29(4):321–330, 1997.

[12] A. W. Fitzgibbon and A. Zisserman. Automatic 3D model acquisition and generation of new images from video sequences. In *Proc. of European Signal Processing Conf. (EUSIPCO '98), Rhodes, Greece*, pages 1261–1269, 1998.

[13] Institute of Industrial Science(IIS), The Univ. of Tokyo. *Urban Multi–Media/3D Mapping workshop*, Japan, 1999.

[14] Digital Michelangelo Project, 2000. http://graphics.Stanford.EDU/projects/mich/.

[15] M. Reed and P. K. Allen. 3-D modeling from range imagery. *Image and Vision Computing*, 17(1):99–111, February 1999.

[16] V. Sequiera, K. Ng, E. Wolfart, J. Concalves, and D. Hogg. Automated reconstruction of 3D models from real environments. *ISPRS Journal of Photogrammetry & Remote Sensing*, 54:1–22, 1999.

[17] H.-Y. Shum, M. Han, and R. Szeliski. Interactive construction of 3D models from panoramic mosaics. In *IEEE Conf. Computer Vision and Pattern Recognition*, Santa Barbara, CA, June 1998.

[18] I. Stamos. *Geometry and Texture Recovery of Scenes of Large Scale.* PhD thesis, Columbia University, 2001.

[19] I. Stamos and P. K. Allen. 3–D model construction using range and image data. In *IEEE Conf. Computer Vision and Pattern Recognition*, volume I, pages 531–536, Hilton Head, SC, July 2000.

[20] I. Stamos and P. K. Allen. Registration of 3D with 2D imagery in urban environments. In *Inter. Conf. on Computer Vision*, Vancouver, Canada, July 2001.

[21] MIT City Scanning Project, 2000. http://graphics.lcs.mit.edu/city/city.html.

[22] Visual Information Technology Group, Canada, 2000. http://www.vit.iit.nrc.ca/VIT.html.

[23] H. Zhao and R. Shibasaki. A system for reconstructing urban 3D objects using ground–based range and CCD sensors. In *Urban Multi–Media/3D Mapping workshop*, Inst. of Industr. Sc., The Univ. of Tokyo, 1999.

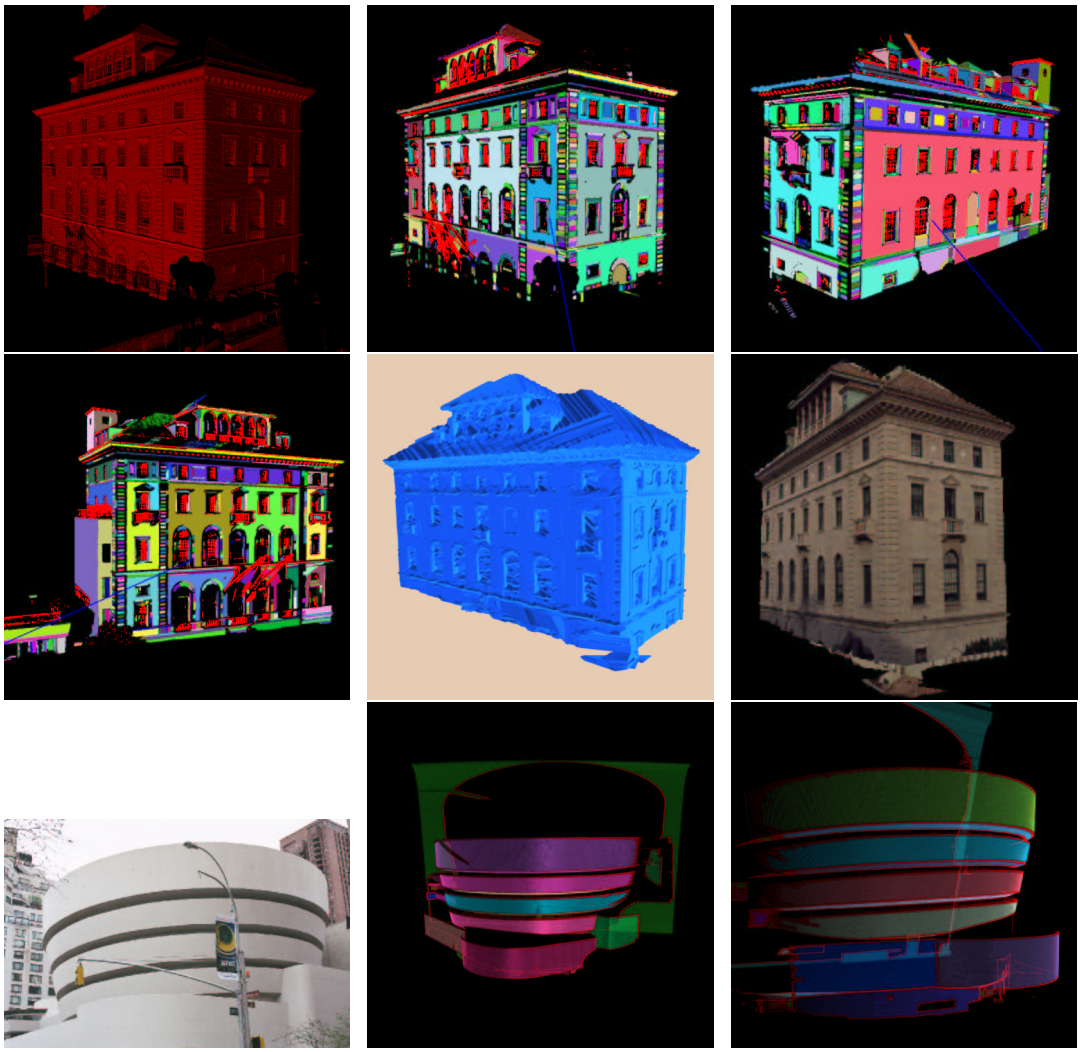

Figure 2: **Top Row:** Casa Italiana. a) Range-scan of first view of the building (1 million points). b,c) Segmented scans of first view and second view (each segmented surface is displayed with different color). **Middle Row:** Casa Italiana. d) Segmented scan of third view. e) Volumetric solid model of the building. f) Photograph of building texture-mapped on the solid model. The registration between the photograph and the model is automatic. **Bottom Row:** Guggenheim Museum, New York City. g) Photograph of the building. e,f) Two segmented scans of the building.

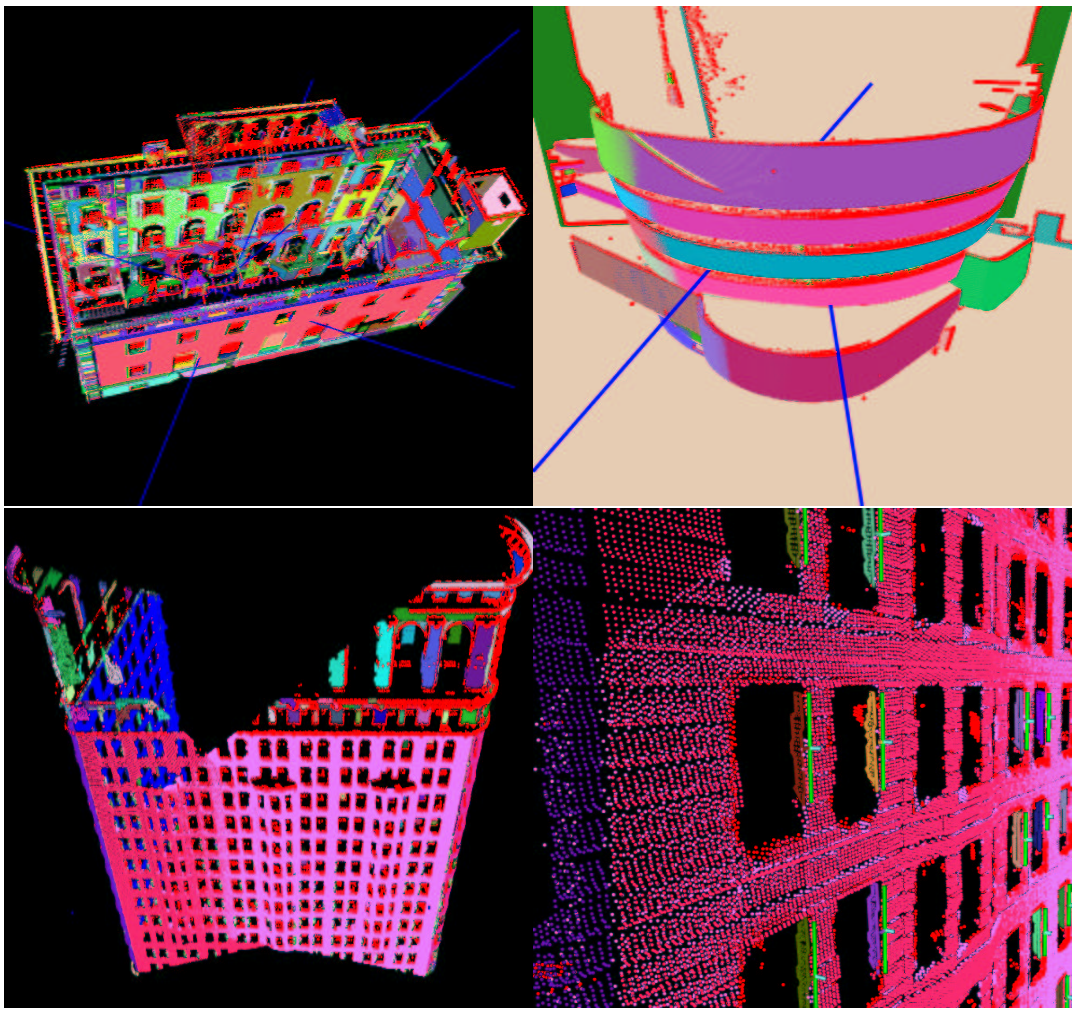

Figure 3: Registration of a) 3 range scans of the Casa Italiana, b) 2 range scans of the Guggenheim Museum and c) 2 range scans of the Flat Iron Building. d) Close view of registration of Flat Iron Building.

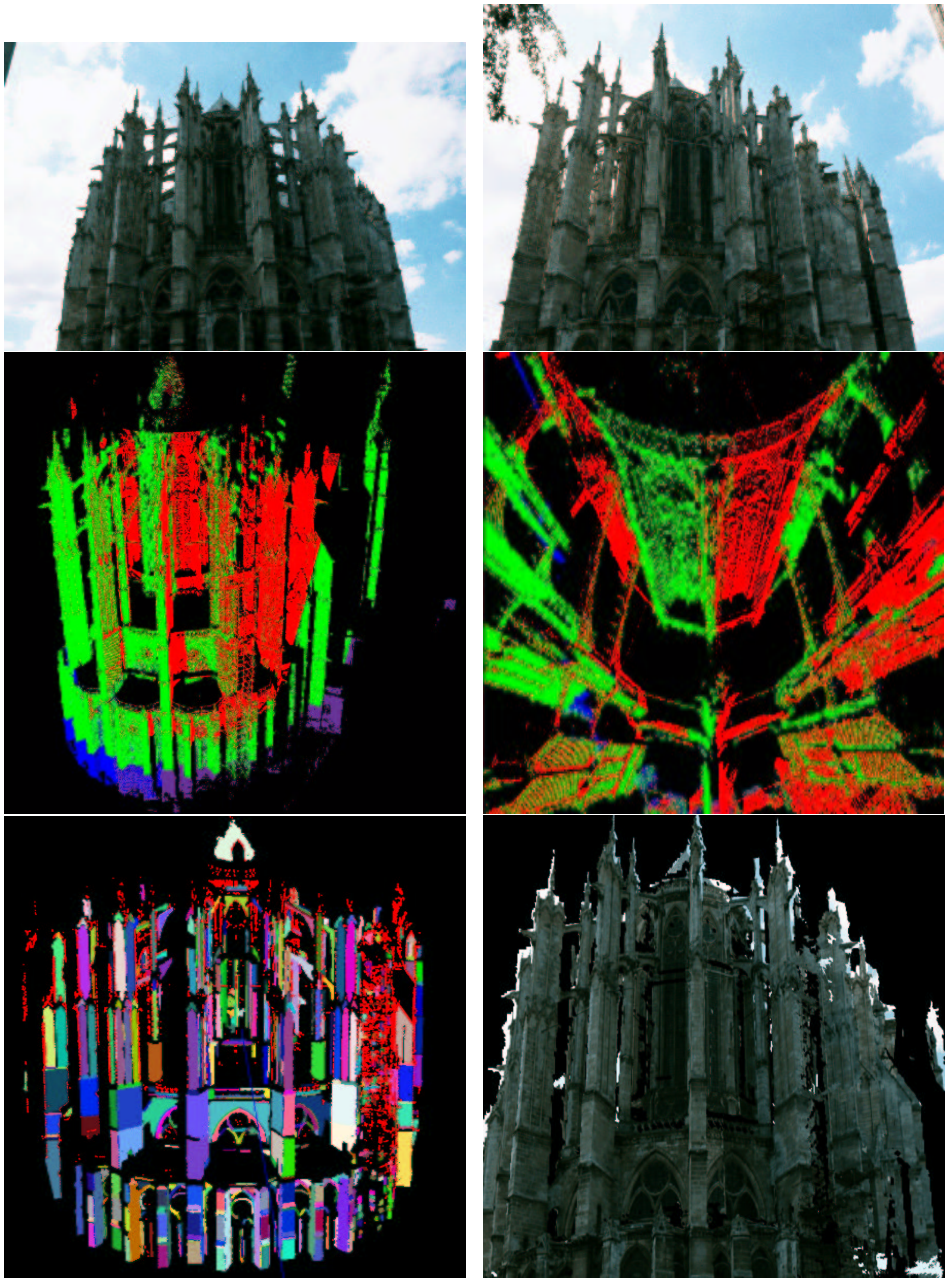

Figure 5: **First Row:** a,b) Two images of the exterior of the St. Pierre Cathedral. **Second Row:** c) Four registered range scans (each different scan is represented with a different color). The registration is the result of matching automatically extracted features between views. d) Registration detail. **Third Row:** e) Segmentation of one range scan (different surfaces are displayed with different color). f) Texture-map of the two photographs a) and b) on the 3-D model of the four range scans.