ImagiCAD: Experimental Image Based Modeling System

Victor Lempitsky*
Moscow State University

Denis Ivanov†
Moscow State University
RL Labs JSC

Anton Shokurov‡
Moscow State University

Yevgeniy Fedotov§
Moscow State University

Yevgeniy Kuzmin¶
Moscow State University
RL Labs JSC

Figure 1: A photograph and a model acquired with ImagiCAD

Abstract

Image based approach to modeling has recently emerged as an attractive alternative to the traditional modeling methods. A number of research and commercial image based modeling systems has appeared, employing novel results in the computer vision field. Most of these systems are targeted at acquisition of highly-realistic virtual models. Other applications, such as non-intrusive metrology, are also considered.

The paper is dedicated to an image based modeling system developed at Moscow State University and RL Labs JSC. We briefly overview the core computer vision algorithms employed in the system. The pecularities of the user interface and the results obtained with the system are presented in the end of the paper.

Keywords: image based modeling, structure and motion, multiview reconstruction, model acquisition, image based metrology

1 Introduction

Image based modeling has recently become one of the most important computer vision applications. In particular, the problem of model acquisition from 2D photographs have been investigated in details (e.g., [Debevec et al. 1996; Dedieu et al. 2001]). Several commercial systems capable of such interactive modeling are now available [Canoma ; Photomodeler ; RealViz] (for the review and comparison see [Dedieu et al. 2001]).

These systems share a lot in common. Most of them rely on intensive user input in order to establish correspondences between input images. These correspondences are processed and the parameters of the cameras used to take the photographs (*motion*) and the 3D locations of scene elements (*structure*) are estimated. Thus, these systems employ a class of algorithms, usually called *structure-from-motion* (or, more precisely, *structure-and-motion*, since both components are estimated simultaneously and neither of both is prior to another).

Structure-and-motion algoritms and, more generally, multiview geometry algorithms have been intensively investigated within computer vision community for a long period of time. The results obtained are collected and thoroughly discussed in an excellent textbook [Hartley and Zisserman 2000]. Besides geometric issues, to produce good results such systems must perform an efficient optimization of a non-linear functional, corresponding to reprojection error (so-called *bundle adjustment*). For a good review on bundle adjustment see [Triggs et al. 2000].

When 3D scene structure and motion are estimated, it becomes possible to use it for various modeling tasks. In particular, it is possible to reconstruct either with small user intervention (as implemented in most systems) or even without it (the subject of ongoing research, e.g., [Dyer 2001]) a three-dimensional model of a scene. Such models are usually textured with the texture maps extracted directly from the photographs and therefore can be considered truly photo-realistic. It usually takes much more time for a designer to create a model of comparable quality using traditional modeling tools.

*e-mail: vitya@fit.com.ru
†e-mail: dvi@fit.com.ru
‡e-mail: anton@fit.com.ru
§e-mail: efedotov@fit.com.ru
¶e-mail: yekuzmin@fit.com.ru

Non-intrusive metrology is another application of great importance. Basing on reconstructed structure, one can interactively measure the distances and angles between 3D elements of a scene. This technology is the most beneficial when direct measurements of the model are impossible or highly undesired.

In this paper, we present an image based modeling system called ImagiCAD (Figure 1). It can be regarded as a typical image based modeling system of a kind. In it, the modeling process starts with manual correspondence selection. The user uploads the photographs of the object, and then selects the corresponding projections of object point features. He/she can also select the projections of straight lines and impose incidence, coplanarity, and parallism constraints on the spatial arrangement of the features.

When feature projections are selected, structure and motion recovery process is performed. As a result, the estimates for feature locations (structure) and for camera parameters (motion) are obtained. The recovered feature locations can be used for interactive construction of a virtual model or for image-based measurements.

While developing the system, we have investigated and tested a bunch of structure-and-motion algorithms. We briefly report on these investigations and present the resulting estimation algorithm in section 2. Section 3 presents the most interesting details of our correspondence selection interface as well as the model construction and metrology processes along with their sample results.

2 Structure and Motion Recovery

In this section we briefly overview the approaches to structure and motion reconstruction implemented in our system. All the algorithms mentioned below are discussed in details in [Hartley and Zisserman 2000].

2.1 Projective reconstruction

Projective geometry provides a useful and efficient framework for the reconstruction process. Furthermore, in the absence of camera calibration, structure and motion parameters can be estimated only up to a projective reconstruction of the world 3D space. It is now realized that even in the presence of camera calibration projective algorithms are much more stable than those invoking operations with calibrated cameras in euclidean space. Therefore, projective calibration is now regarded as a necessary step in any structure-and-motion reconstruction process with the upgrade to metric geometry to follow (a matter that we touch in section 2.3). Below we briefly overview the way how the projective reconstruction is performed in our system.

We assume that we are given with N 2D images $I_1, I_2 \dots I_N$. The structure to be recovered consists of S 3D points $M^1, M^2 \dots M^S$ represented with homogeneous 4-vectors. User-selected correspondences are presented as an incomplete set of 2D projections m^i_j, where m^i_j denotes the projection of a point M^i on the image I_j represented with homogeneous 3-vector $(x, y, 1)$, where x and y are the projection coordinates in the properly normalized coordinate system of the image.

The image formation process for the image I_j is modeled as

$$m \cong \mathbf{P_j} M, \tag{1}$$

where M is a homogeneous 4-vector, representing a point in 3D space, m is a homogeneous 3-vector, representing its 2D projection on the image I_j, $\mathbf{P_j}$ is a 3×4 projection matrix representing internal and external camera parameters, and $\cong$ denotes equality up to scale. Equation (1) is a generalization of a well known pinhole camera model. At the same time, it is not able to model radial distortion, and the significant presence of this effect requires the lens to be precalibrated and the images to be prelimenary undistorted.

Thus our problem is formalized as follows. Given a set of projections $\{m^i_j | (i,j) \in \Omega \subset 1 \dots S \otimes 1 \dots N\}$ estimate $\mathbf{P_1}, \mathbf{P_2} \dots \mathbf{P_N}$ and $M^1, M^2 \dots M^S$, such that reprojections $\mathbf{P_j} M_i$ lie as close as possible to given projections m^i_j. The notion of closeness between $\mathbf{P_j} M_i$ and m^i_j can be formalized in two ways via *algebraic* or *geometric* reprojection errors:

$$\begin{aligned} {\rho_{alg}}^i_j = ((\mathbf{P_j} M_i)[1] \cdot m^i_j[3] - (\mathbf{P_j} M_i)[3] \cdot m^i_j[1])^2 + \\ ((\mathbf{P_j} M_i)[2] \cdot m^i_j[3] - (\mathbf{P_j} M_i)[3] \cdot m^i_j[2])^2 \end{aligned} \tag{2}$$

$${\rho_{geom}}^i_j = (\frac{(\mathbf{P_j} M_i)[1]}{(\mathbf{P_j} M_i)[3]} - m^i_j[1])^2 + (\frac{(\mathbf{P_j} M_i)[2]}{(\mathbf{P_j} M_i)[3]} - m^i_j[2])^2 \tag{3}$$

Being quadratic on the coefficients of $\mathbf{P_j}$ and M^i, the algebraic error is more computationally tractable. Thus, the problems

$$\mathbf{P_j} = argmin \sum_{i=i_1 \dots i_k} {\rho_{alg}}^i_j, \quad s.t. \sum_{k=1..12} \mathbf{P_j}[k]^2 = 1 \tag{4}$$

and

$$M^i = argmin \sum_{j=j_1 \dots j_l} {\rho_{alg}}^i_j, \quad s.t. \sum_{k=1..4} M^i[k]^2 = 1 \tag{5}$$

can be solved via SVD decomposition.

However, the algebraic error is dependent on arbitrary scaling of homogeneous vectors, and therefore it is the geometric error that should be chosen as an ultimate measure of reconstruction quality. Our system therefore acts as follows. Reconstruction process starts from the estimation of epipolar geometry for an automatically chosen pair of images. Basing on the epipolar geometry one can recover the projection matrices for each of both images. After these cameras are estimated, the system sequentially adds points and cameras to the reconstruction by solving problems (4) or (5). Each time, the system chooses the next entity to be estimated basing on the singular values of the linear system corresponding to the respective minimization problem.

After this augmentation process, the system performs global optimization of total geometric error:

$$\begin{aligned} \{\mathbf{P}, M\} = argmin \sum_{(i,j) \in \Omega} {\rho_{geom}}^i_j \\ s.t. \qquad (\mathbf{P_j} M_i)[3] > \varepsilon \sum_{k=1..12} \mathbf{P_j}[k]^2 = 1 \sum_{k=1..4} M^i[k]^2 = 1 \end{aligned} \tag{6}$$

This process is usually referred as bundle-adjustment [Triggs et al. 2000]. Various optimization methods can be employed for such minimization, of which Levenberg-Marquardt algorihm is the most widely used. In our system, we employ a more sophisticated algorithm of constrained optimization [Fletcher and Leyffer].

2.2 Constrained reconstruction

Our system is capable of recovering the structure comprising not only points but also lines and planes. Planes in the 3D projective space are represented with homogeneous 4-vectors, and are easily tractable. Lines are much more difficult to deal with, since

the four-dimensional manifold of lines in 3D space has no simple parametrization. In our system, a line is represented with a pair of points in two predefined planes and simultaneously with a pair of planes containing two predefined points.

Our system can reconstruct 3D lines from their user-selected projections on images. This can be accomplished using two-point representation. Each selected projection of a line imposes the constraint that the reprojections of both 3D points lie on the line projection. Algebraically, this is expressed with a pair of simple linear equations.

Structure constraints that can be imposed in our system include all possible incidence constraints (i.e. one entity belongs to/contains another entity, e.g., a line belongs to a plane). Constraints of parallelism can also be easily introduced. To do this, the user creates a logical *direction* and specifies all entities (lines and planes) parallel to this direction. On the algorithmical level, each direction is represented with a point. For lines and planes parallel to the direction, the system adds constraints of incidence to the introduced point. Finally, all points representing directions are constrained to lie on a specially introduced plane (so-called *plane at infinity*). On the stage of projective reconstruction, direction points and plane at infinity are treated like other points and planes. Besides user defined constraints, we impose incidence constraints between points and planes representing the same line.

All incidence constraints can be expressed with equations linear on the coefficients of each participating entity. Minimizing the quadratic error of the equation set corresponding to one entity subject to normalization constraint lead to the computational problem similar to (4) and (5) [1]. Therefore, points and lines can be estimated in the augmentation process. Lately, lines and planes along with all constraints are introduced into the bundle adjustment process.

In our experiments, we found that constraints can significantly increase reconstruction quality. For some scenes (e.g. complex intéreurs), reconstruction without constraints is impossible, since in the presence of minor selection imprecisions point-based bundle-adjustment (6) converges to the incorrect state due to the error accumulated during augmentation stage.

2.3 Metric reconstruction

So far, we discussed the problem of projective structure-and-motion reconstruction, e.g. reconstruction up to an arbitrary projective transformation. Such freedom is however totally unacceptable for most applications. To fix the problem and to obtain metric reconstruction (i.e. reconstruction up to scale, rotation, and translation), one must employ additional knowledge about either structure or motion. In the structure domain, our system is capable to use the reconstructed plane at infinity to constrain the reconstruction up to an arbitrary affine transformation. If three points corresponding to orthogonal directions are known than it is possible to upgrade the reconstruction to metric state.

In recent years, a great amount of research work has been dedicated to metric upgrade using additional knowledge about internal parameters of cameras (so-called *auto-* or *self-calibration*). Ability of such upgrade can be illustrated as follows. Let $\mathbf{P_1}, \mathbf{P_2} \dots \mathbf{P_N}$ be a set of reconstructed projection matrices. Since without prior knowledge our reconstruction is defined up to a projective transforamtion, our set is equivalent to a set $\mathbf{P_1}T, \mathbf{P_2}T \dots \mathbf{P_N}T$ where T is an arbitrary non-singular 4×4 matrix. Using the theorem of QR-decomposition any projection matrix can be decomposed as

$$\mathbf{P_j} = K_j(R_j|-R_jt_j), \tag{7}$$

where K_j is upper-triangular, R_j is a rotation matrix, and t_j is a 3-vector. For a metric camera these three parts have particular meanings, K_j being a matrix of internal camera parameters, R_j being a matrix defining camera orientation, and t_j being a position of the camera viewpoint in metric space. Therefore, when some of the internal parameter matrices are (partially) known, it is possible to choose a particular transformation T such that corresponding set of cameras would yield the most correct set of K_j. This transformation will bring the system as close to the metric state as possible.

Though several algorithms exist that are able to perform self-calibration even from the knowledge of pixel rectangularity (corresponding to a particular zero coefficient in matrices K_j), robust self-calibration algorithm even from a reacher knowledge is still an open problem. Our implementation of some self-calibration algorithms does not always yield correct result for noisy data even when the whole matrix of internal camera parameters is known for each camera.

To accomplish metric reconstruction without performing self-calibration, we can also perform reconstruction using non-projective algorithms that directly yields metric reconstruction [2]. Such algorithms have been being long out of favour due to their low robustness against noise. To filter out the noise in the user input, we start with the projective reconstruction and then replace original user selected projections with reprojections, resulted from estimated projective reconstruction. We performed multiple experiments with artificial and real data and verified that this simple scheme is very stable and yields result even when initial data are plagued by significant noise. In particular, we use motion estimation via essential matrices to estimate motion for different pairs of images. Such pairwise reconstructions are combined using unifying metric transformations, estimated from relative positions of common points.

Whatever is the chosen strategy for obtaining metric reconstruction, the concluding step is always a metric bundle adjustment process:

$$\begin{aligned} \{\mathbf{P}, M\} \quad &= \quad argmin \sum_{(i,j)\in\Omega} {\rho_{geom}}_j^i + \lambda \|\hat{\mathbf{P}}_\mathbf{j}\hat{\mathbf{P}}_\mathbf{j}^T - K_jK_j^T\|^2 \\ s.t. \qquad & (\mathbf{P_j}M_i)[3] > \varepsilon \qquad \sum_{k=1..4} M^i[k]^2 = 1, \end{aligned} \tag{8}$$

[1] E.g., for a point ${}_{L_k}M'$ from a pair of points representing line L_k we get the following problem

$$\begin{aligned} {}_{L_k}M' = \underset{M}{argmin}(&\omega_1(({}_{L_k}\Pi' \cdot M)^2 + ({}_{L_k}\Pi'' \cdot M)^2) + \\ &\omega_2({}_{L_k}\hat{\Pi}' \cdot M)^2 + \omega_3 \sum_{r=r_1\dots r_k} (\Pi_r \cdot M)^2 + \\ &\omega_4 \sum_{j=j_1\dots j_l} (\mathbf{P_j}M \cdot l_j^k)^2), \quad s.t. \sum_{k=1..4} M[k]^2 = 1 \end{aligned}$$

where ${}_{L_k}\Pi'$ and ${}_{L_k}\Pi''$ are two planes defining the same line, ${}_{L_k}\hat{\Pi}'$ is a plane constraining the point, Π_r are the planes incident to the line, $\mathbf{P_j}$ are the projection matrices for the images containing projections of the line, l_j^k are these projections, ω_i are predefined weighting coefficients, and $(\cdot)$ denotes scalar product of 4- or 3-vectors.

[2] These algorithms assume that internal camera parameters are known. This assumption is not too limiting for many applications. Effective camera calibration software can be easily assembled using e.g. [OpenCV]. Indeed, camera precalibration is often a must even for projective algorithms due to the presence of radial distortion.

where $\hat{\mathbf{P}}_\mathbf{j}$ is a left 3×3 matrix of $\mathbf{P_j}$, K_j is a matrix of internal camera parameters, and λ is a weighting factor. For constrained reconstruction the first term should be augmented with the terms representing constraints and line reprojection errors. The second term measures the difference between the first factor in decomposition (7) and the given internal camera parameters matrix. If the internal parameters are only partially known, then the norm in the second term should comprise squared differences over given coefficients.

3 Interface and Applications

3.1 User interface

As image based modeling systems usually require significant user interaction, convinient user interface is the crucial part of such systems. In our system, the main burden is selecting multiple point correspondences. Therefore, we put many efforts in the creation of effective point selection tool. Besides convenient zooming and paning mechanism, we implement an automatic user guiding procedure, based on multiview geometry. This gudance becomes active when the user have selected the projections of a point on two images. Starting from the third image the system will show to the user an approximate position of point projection deduced from multiple view constraint [3].

Another useful tool for point selection is the *point audit*. This routine is aimed at detection of significant errors in the user input (e.g., the situations when the user was occasionally confused with some repetitive pattern in the scene and selected a projection of analogeous point instead of the proper one). Point audit employs the RANSAC algorithm [Torr et al. 1994] to estimate multiview geometry [4]. The RANSAC algorithm allows to estimate multiview geometry even in the presence of correspondences selected with significant errors (*outliers*). Furthermore, when multiview geometry relations are estimated, these erroneous correspondences can be distinguished and reported to the user.

In our experiments, we found out that the described interface improvements can significantly speed up the correspondence selection process. This makes our system a useful tool for various applications of image based modeling.

3.2 Model acquisition

The first application to be mentioned is the acquisition of 3D models. The models produced with our system comprise surface patches and 3D polylines.

Once structure and motion estimation has been performed, the model can be composed interactively. To add a patch to the model, the user selects a set of reconstructed points $M^{i_1}, M^{i_2}, \ldots M^{i_K}$ by clicking on their projections on one of the views. This set of points is triangulated and the triangles obtained are textured with the corresponding image region. The resulting patch is added to the model.

To add a 3D curve to the reconstruction, the user selects its ends that should be also reconstructed points M^{i_1} and M^{i_2}. The user also selects several projections of the curve in a form of 2D polylines. After that, the curve is reconstructed as a polyline with a predefined number of vertices via iterative reprojection energy minimization.

The resulting models (Figure 5) can be rendered with the internal OpenGL-based viewer [OpenGL] of the system. Export in VRML format [Carey and Bell 1997] is also available, making it possible to use the models in a multitude of virtual reality applications (games, virtual tourism, virtual heritage). Since, the camera parameters are recovered along with the scene structure one may consider using reconstructions in augmented reality tasks.

3.3 Image based metrology

Our system can also be used as a metrology tool. The structure recovered after metric reconstruction is visualized, allowing the user to measure interactively the distances and angles between reconstructed 3D elements. Very high metrology precision with tenths of percent order of relative errors can be attained for properly calibrated cameras and an accurate user input. This makes our system usable for various control tasks.

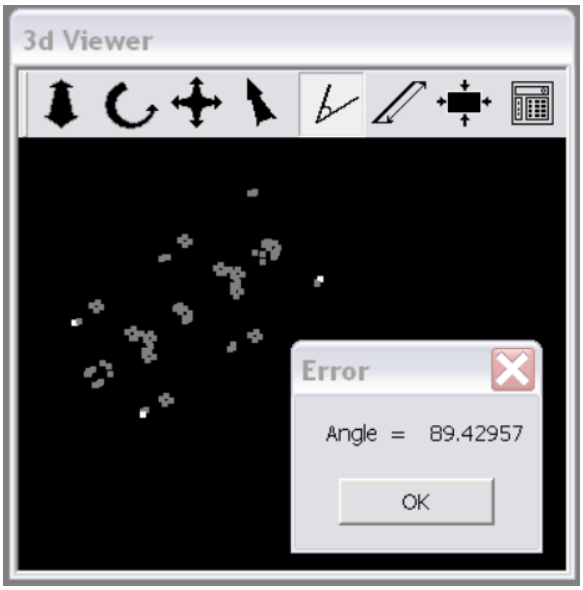

Figure 2: Metrology session in ImagiCAD. The user has measured one of the angles on the reconstructed pylon.

We are currently investigating the possibility of applying our system to the control of deformation of industrial objects (in particular, electric transmission line pylons). We are planning to develop a software tool specialized to this particular task on the base of ImagiCAD.

4 Conclusion

We have presented an image based modeling system. It exploits the wide range of the recently emerged algorithms and demonstrates the broad capabilities of image based modeling in such tasks as model acquisition and metrology. It also possesses a covenient user interface, which provides the user an intellectual aid during the correspondence selection process.

The input for our system consists of several photo images, taken with an off-the-shelves digital still camera. On these images, the

[3] We estimate a trifocal tensor (see [Hartley and Zisserman 2000]) for a corresponding triple of images. A trifocal tensor is a $3 \times 3 \times 3$ tensor, allowing for computation of a point projection on a third view given its projections on the other two views.

[4] Here we again use trifocal tensors for different triples of views.

user select corresponding projections of an object's feature points. Line projections can be selected as well. Moreover, the user can point out the relations of incidence, parallelism or coplanarity for the points and lines under reconstruction.

Selected feature projections along with imposed constraints are invoked into structure and motion recovery process. It starts with the projective uncalibrated reconstruction and as a second step recovers the metric structure of the scene. To do the latter step, it requires at least partial knowledge about internal camera parameters.

In our system, the reconstructed 3D elements can be employed either for model construction or for metrology. To construct the model of the object the user interactively selects the set of points on any image. This set is triangulated and the obtained triangles textured with the corresponding image regions are added to the model. The model can also be augmented with the spatial curves, represented as polylines and reconstructed from user selected projections.

The second important application is the image based metrology. The user can interactively measure the distances and the angles between reconstructed 3D elements. Very high metrology precision with tenths of percent order of relative errors can be attained.

A lot can be added to our system. In fact, it can be regarded as a platform for more specific image based modeling applications. Possible improvements include more sophisticated scene structure model, specialization of user interface to some particular tasks, automatization of correspondence selection and model acquisition processes.

5 Acknowledgements

This work is supported in part by the grant of the Russian Foundation for Basic Research # 03-07-90381.

References

CANOMA. http://www.canoma.com.

CAREY, R., AND BELL, G. 1997. *The Annotated VRML 2.0 Reference Manual*. Addison-Wesley.

DEBEVEC, P., TAYLOR, C., AND MALIK, J. 1996. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In *Proceedings of SIGGRAPH 1996*, ACM Press / ACM SIGGRAPH, Computer Graphics Proceedings, Annual Conference Series, ACM, 11–20.

DEDIEU, S., GUITTON, P., SCHLICK, C., AND REUTER, P. 2001. Reality: an interactive reconstruction tool of 3d objects from photographs. In *In Proceedings of Vision Modeling and Visualization'2001*, 195–202.

DYER, C. R. 2001. Volumetric scene reconstruction from multiple views. In *Foundations of Image Understanding*, L. S. Davis, Ed. Kluwer, 469–489.

FLETCHER, R., AND LEYFFER, S. Nonlinear programming without a penalty function. Numerical Analysis Report NA/171.

HARTLEY, R., AND ZISSERMAN, A. 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

OPENCV. Intel library. http:// www.sourceforge.net/ projects/ opencvlibrary.

OPENGL. http://www.opengl.org.

PHOTOMODELER. http://www.photomodeler.com.

REALVIZ. http://www.realviz.com.

TORR, P., BEARDSLEY, P., AND MURRAY, D. 1994. Robust vision. In *Proceedings British Machine Vision Conference*, 145–155.

TRIGGS, B., MCLAUCHLAN, P., HARTLEY, R., AND FITZGIBBON, A. 2000. Bundle adjustment - a modern synthesis. In *Vision Algorithms: Theory and Practice*, Springer-Verlag, LNCS 1883, 298–372.

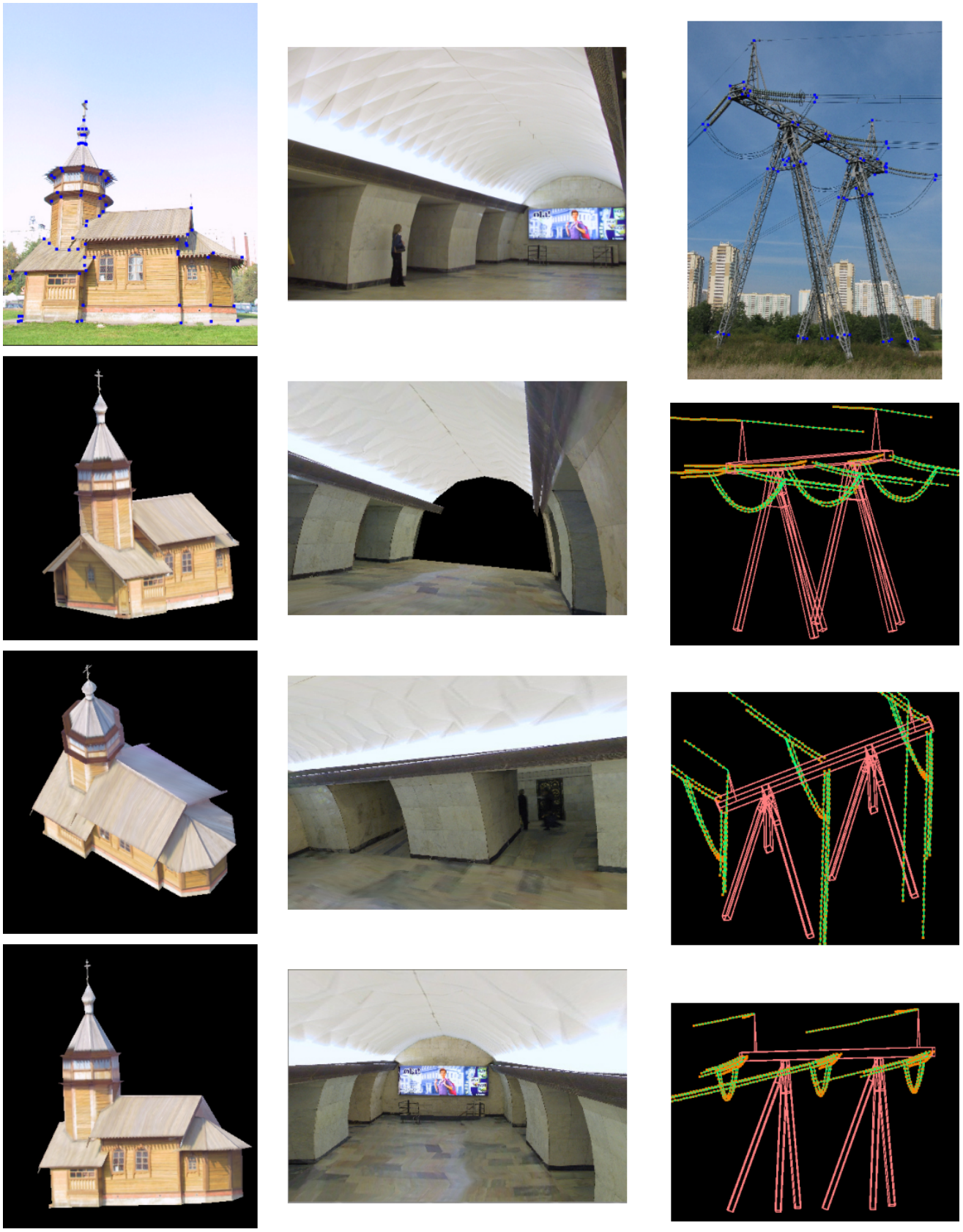

Figure 3: Reconstructed scenes. Top row – one of the intial photographs. Lower rows – reconstructed models rendered from different viewpoints.