3D RECONSTRUCTION AND VISUALIZATION OF URBAN SCENES FROM UNCALIBRATED WIDE-BASELINE IMAGE SEQUENCES

Helmut Mayer

Institute of Photogrammetry and Cartography, Bundeswehr University Munich, D-85577 Neubiberg, Germany
Helmut.Mayer@unibw.de

KEY WORDS: 3D Reconstruction, Auto-Calibration, Markerless Orientation, Visualization

ABSTRACT:

In this paper we show how based on a number of different techniques it is possible to fully automatically generate basic ingredients for high quality visualizations of urban areas characterized by vertical facade planes from nothing but uncalibrated wide-baseline image sequences without using any markers or ground control. At the core of our algorithms are least-squares matching, projective geometry based reconstruction, robust estimation based on random sample consensus – RANSAC, auto-calibration, projective and Euclidean bundle adjustment, plane to plane homographies, as well as the robust estimation of image mosaics. Results for the Hradschin in Prague, Czechia, and Plaza Real in Barcelona, Spain, show the potential and shortcomings of the employed algorithms.

1. INTRODUCTION

Recent years have seen a couple of approaches for the fully automatic generation of three-dimensional (3D) Euclidean models from uncalibrated image sequences, among the most advanced of which are (Pollefeys, Van Gool, Vergauwen, Verbiest, Cornelis, and Tops, 2004) and (Lhuillier and Quan, 2005). The approaches usually consist of the robust estimation of a projective reconstruction and n-fold correspondences followed by calibration and possibly dense depth estimation, all usually restricted to small images with a short baseline, e.g., from a video camera.

Opposed to this, we aim at applications where higher image resolutions in the range of several Megapixels are given as input, obtained, e.g., from consumer digital cameras in the range of several hundred US $. Because of the lower frame rates (one image can usually be taken on a sustained basis about every second on average) and higher data volumes per image it is natural to take images with a wider baseline making the matching of points between the images severely more difficult. Therefore, we show how employing high precision to become more reliable it is possible to obtain 3D reconstructions of rather difficult scenes with many occlusions and partly close to no 3D structure.

The focus of this paper is on urban scenes. Therefore, it is reasonable to use at least partly for the modeling and visualization of the scenes planes, particularly the vertical planes of the facades. As larger parts of our scenes are assumed to be captured in at least three images, it becomes on one hand necessary to fuse the information from the individual images on the detected planes. Yet, on the other hand, it gives us the opportunity, to separate by means of consensus between pixels taken from different images the information on the plane from off-plane information, allowing us to generate a "cleaned" version of the image on the plane without many of the occlusions in the individual images. This part has been inspired by (Wang, Totaro, Taillandier, Hanson, and Teller, 2002) and (Böhm, 2004). Yet, opposed to the latter, we fully automatically and robustly generate the planes and from them the two-dimensional (2D) homographies, i.e., plane to plane mappings. We also integrate the planes into our 3D model and generate visualizations from it.

Impressive results in terms of visualization of urban scenes have been shown by (Debevec, Taylor, and Malik, 1996) by taking the image from the (real) camera closest to the current (virtual) viewpoint, though the 3D model employed has been generated manually. Then, there is work for architectural scenes which goes far beyond what we are presenting here in the sense that much more knowledge about the structures and regularities of urban scenes is used. The most sophisticated example today is probably (Dick, Torr, and Cipolla, 2004) employing a statistical generative model based on Markov Chain Monte Carlo (MCMC) sampling. Closer to our work as it is more geometry-based is (Werner and Zisserman, 2002). Yet, compared to our work they employ perpendicular vanishing points for auto-calibration and 3D reasoning which restricts the work as given to scenes with three perpendicular main directions. They have also only shown results for image triplets.

In the remainder of this paper, we first present our approach for 3D reconstruction from wide-baseline image sequences (cf. Section 2.). The obtained 3D Euclidean model is the basis for deriving vertical facade planes. For them facade images at least partly "cleaned" from occlusions are computed by means of median or consensus between the pixels projected onto the planes from different camera positions (cf. Section 3.). In Section 4. we present additional results and we end up with conclusions.

2. 3D RECONSTRUCTION

Our approach is aiming at wide-baseline image sequences made up of images of several Megapixels. We make the following assumptions for 3D reconstruction:

- The camera constant (principal distance) is constant. Yet this is not as restrictive as it may sound because we found that the influence of auto-focusing that one cannot switch off for some cameras we use can mostly be neglected for the distances typical for urban applications. We also assume that the principal point is close to the image center. This is the case for practically all digital cameras, and would only not hold if parts of images were taken.
- The images are expected in the form of a sequence, with at least three-fold overlap for all images.
- The camera is not to be rotated around the axis of the objective between consecutive images, though rotations below $\pm 20^\circ$ usually do not degrade the result considerably.

Our basic idea to obtain a reliable result is to strive for a very high precision in the range of 0.05 to 0.3 pixels by means of least-squares matching and bundle adjustment. If this value is higher

or lower depends in first instance on scene geometry and geometrical quality / stability of the camera, but in second instance also on lighting conditions, etc. The overall reasoning is that it is (extremely) unlikely that a larger number of non-homologous points conspire to achieve a highly precise result by chance.

Based on this idea we start using Förstner points (Förstner and Gülch, 1987). They are matched via cross-correlation. In color images the coefficient for the channel where the variance is maximum is taken. To avoid multiple, i.e., indecisive matches (e.g., upper right corners of windows on larger facades can look very similar), we match the result in the search image back into the given image and only accept it, if the original position is found to be the maximum again. Point pairs checked via correlation are then refined via least-squares matching with an affine geometrical model. The latter is also used for three- and more-fold images. In all cases we compute the complete covariance information.

The highly precise points are the basis for a projective reconstruction employing fundamental matrices $\mathbf{F}$ and trifocal tensors $\mathcal{T}$ (Hartley and Zisserman, 2003). If calibration information is available, we use (Nistér, 2004) to determine the Euclidean 3D structure for image pairs. As in spite of our efforts to obtain reliable matches we obtain partly less than 20% of correct homologous points for difficult scenes, we employ Random Sample Consensus – RANSAC (Fischler and Bolles, 1981) for the estimation of $\mathbf{F}$ and $\mathcal{T}$. Because we do not only have rather low numbers of correct matches (inliers), but as these inliers are also partly very unevenly distributed over the image and thus not all of them lead to a correct model, i.e., a model representing all inliers with the inherent, yet unknown geometric precision, we employ a variant of the locally optimized RANSAC scheme of (Chum, Matas, and Kittler, 2003). While they take a larger number, i.e., 50%, of random samples from the maximum set of inliers derived at a certain stage to derive an improved estimate, we take the whole maximum set and employ robust bundle adjustment (Hartley and Zisserman, 2003; Mikhail, Bethel, and McGlone, 2001). The latter is done for two iterations, always using the outcome of the bundle adjustment to derive new sets of inliers.

The employed bundle adjustment is suitable for the projective as well as the Euclidean case. We model radial distortion with a cubic and a quartic term. Bundle adjustment takes into account the full covariance information derived by least-squares matching. We estimate the precision of the residuals and use them in two ways to make the adjustment robust: First, we reweight the observations based on the ratio of the size of the residual and its variance. Second, after convergence we throw out all points with a ratio beyond three, a value found empirically.

As our images are in the range of several up to possibly tens of Megapixels, it is important to initially restrict the search space for matching. Yet, because we do not want to restrict the user more than given in the assumptions at the begin of the section, we cannot assume that the movement is only vertically or horizontally or that it is even in a certain range. Particularly for urban scenes with very close and far away objects disparities can be rather large, in the extreme case exceeding the image size. We thus take as initial search space the full image, but reduce the image in a pyramid and do the first search on a pyramid level with a size of approximately 100×100 pixels. Here, full search can be done efficiently. Matching and projective reconstruction lead to fundamental matrices and thus epipolar lines on the highest level, restricting the search on the next level considerably. Once trifocal tensors have been determined, the search space becomes a small area in the third image. Trifocal tensors are computed for the second highest level in all cases and additionally on the third highest level if the image size exceeds one Megapixel.

To orient whole sequences, we link triplets based on 3D homographies computed from projection matrices for images common between triplets. (E.g., the triplets $(1, 2, 3)$ and $(2, 3, 4)$ have the images 2 and 3 in common.) Additionally, we project already known 3D points into the newly linked image to generate $i + 1$-fold points, with i being the current number of images a point is visible in. After these steps we bundle adjust the sequence. Once all projection matrices and 3D points have been computed, we track the points generated on the second or third highest level of the pyramid down to the original resolution again via least-squares matching in all images.

If no calibration information has been given, we auto-calibrate the camera employing the approach proposed in (Pollefeys, Van Gool, Vergauwen, Verbiest, Cornelis, and Tops, 2004). It constrains the solution to reasonable values for the parameters, e.g., the principal point corresponds to the center of the image and the camera constant is somewhere in-between one third and three. Auto-calibration is done only once a high quality projective reconstruction has been obtained on the original resolution via projective bundle adjustment. We found that the latter is mandatory, as lower precisions lead to incoherent implicit calibrations of the projective reconstructions, often leading to unacceptable results. Finally, we employ Euclidean bundle adjustment to obtain a highly-precise calibrated 3D model consisting of points and projection matrices including full covariance information.

An example is given in Figures 1 and 2 showing a part of the Hradschin in Prague, Czechia. The back-projection error of the calibrated bundle is $\sigma_0 = 0.16$ pixels in the given 2 Megapixel images and several hundred six-fold points have been computed. One can see that the right angles in the center of the building have been derived very accurately.

Figure 1. Six images of the Hradschin in Prague, Czechia

3. PLANES AND IMAGES ON PLANES

Having obtained a 3D Euclidean model, we assume that an urban scene consists of a considerable number of vertical lines. We can thus orient the model vertically based on the vertical vanishing point derived from the vertical lines and the given calibration information. The vertical vanishing point is detected robustly again using RANSAC, the user only providing the information if the camera has been been very approximately held horizontally or vertically, thus, avoiding to mix up the vertical with a horizontal vanishing point. After detecting the vanishing point, we polish it by means of least-squares adjustment. To make the computation of the vertical direction more robust, we compute vanishing points for a couple, usually if possible five images, derive from all of them the vertical direction of the whole model employing the known rotation of the individual camera, and then finally take the medians in x- and y-direction as the vertical direction.

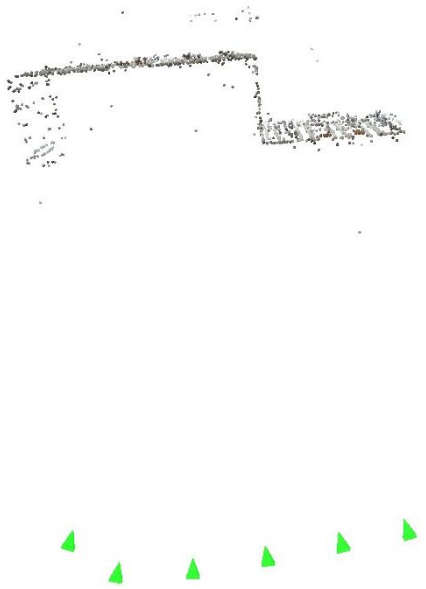

Figure 2. 3D points (red) and cameras (green pyramids, the tip symbolizing the projection center and the base giving the direction of the camera) derived from the images given in Figure 1

The vertically oriented model is the basis for the determination of vertical facade planes using once again RANSAC. For this step one threshold defining the maximum allowed distance of points from the plane has to be given by the user. This is due to the fact that we could determine meaningful thresholds for approximating planes from the covariance matrices via model selection, but this would only take into account the measurement accuracy and not the semantically important construction precision of facade planes.

To make it more robust and precise, we employ the covariance information of the 3D points computed by bundle adjustment by not counting the number of inliers as for standard RANSAC, but testing the distances to a hypothesized plane based on the geometric robust information criterion – GRIC (Torr, 1997). Additionally, we check, if the planes are at least approximately vertical and we allow only a limited overlap of about five percent between the planes. The latter is needed, because of points situated on intersection lines between planes.

From the parameters for the facade planes as well as the projection matrices we compute homographies between the planes and the images. A mapping by a homography **H** between homologous points **x** and **x'** in homogeneous coordinates on a given plane and the image plane of a camera, respectively, is given by

$$\mathbf{x}' = \mathbf{H}\mathbf{x} \quad . \tag{1}$$

If the camera is parameterized as

$$\boldsymbol{P} = \begin{pmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \end{pmatrix} \tag{2}$$

and the plane, the points lie on, with

$$\pi = (\mathbf{n}^\mathsf{T}, d)^\mathsf{T} \quad , \tag{3}$$

and if we parameterize the plane in 2D by setting the Z-component of the plane to zero, **H** is determined as

$$\mathbf{H} = \begin{pmatrix} H_{i1} \\ H_{i2} \\ H_{i3} \end{pmatrix} = \begin{pmatrix} P_{i1} - \pi_1 \cdot P_{i3}/\pi_3 \\ P_{i2} - \pi_2 \cdot P_{i3}/\pi_3 \\ P_{i4} - \pi_4 \cdot P_{i3}/\pi_3 \end{pmatrix} \quad . \tag{4}$$

For the actual mapping of images to a plane one needs to know from which images a plane can be seen from. For it, the information is employed, which 3D points have led to a particular plane, as for the 3D points it is known from which images they were derived. The plane is thought to be visible from the union of the sets of images of all 3D points belonging to a plane. We compute an average image as well as the bias in brightness for each image in comparison to it, also accounting for radial distortion.

The final step is the generation of facade images "cleaned" from artifacts generated by occlusions. The basic information are the projected images normalized via the determined biases in brightness. The cleaning is done by two means, first by sorting the (gray- or color) values and taking the median and second by utilizing the basic idea of (Böhm, 2004). The latter consists in determining an optimum value by means of the consensus between the values for a particular pixel. As (Böhm, 2004) we do not randomly select the values as in RANSAC, but we take the value for a pixel for each image it can be seen from as the estimate and then take as the inliers all, which consent with it. The final result is the average of the inliers.

Results for our running example are given in Figures 3 and 4. From the former one can see that the planes fit nicely to the points. The latter shows the advantages of median and consensus over simple averaging where, e.g., the flag pole at the right hand side is shown several times as a ghost image. The different characteristics of median and consensus are shown more in detail in the additional example in the next section.

4. ADDITIONAL RESULTS

In Figure 8 ten images out of a whole set of 29 uncalibrated images of Plaza Real in Barcelona, Spain, are shown, taken with a Sony P 100 5 Megapixel camera. The basic idea was to walk around the fountain in the center of Plaza Real. From them a 3D model is computed (cf. Figure 5) with $\sigma_0 = 0.18$ pixels after bundle adjustment. As we did not mark our positions when taking the images, the circle around the fountain is more a spiral and could not be closed as the first and last image did not match. Because of this and the fact that larger parts of the facade are planar, it is interesting how well the start (upper left corner) and the end (left side of the sequence) fit together after error accumulation over 29 images. Also the right angles have been determined very well in spite of the relatively large areas where we could not match due to occlusions mostly by the palm trees.

In Figures 6 and 7 additional visualizations are given, in Figure 6 from two actual camera positions (image 11, cf. also Figure 8 upper left image, and 26) and in the second from a position above one of the facade planes. The facade image for the right facade in Figure 5 derived from the ten images given in Figure 8 is given in Figure 9. First, the average image shown at the bottom makes clear by means of the circular streaks how large the influence of radial distortion is for some of the images. (Please note that the images with the largest distortions look from the side onto the plane, strongly amplifying the effect.) Overall, one can see that the average is not acceptable. This is due to the ghost images of

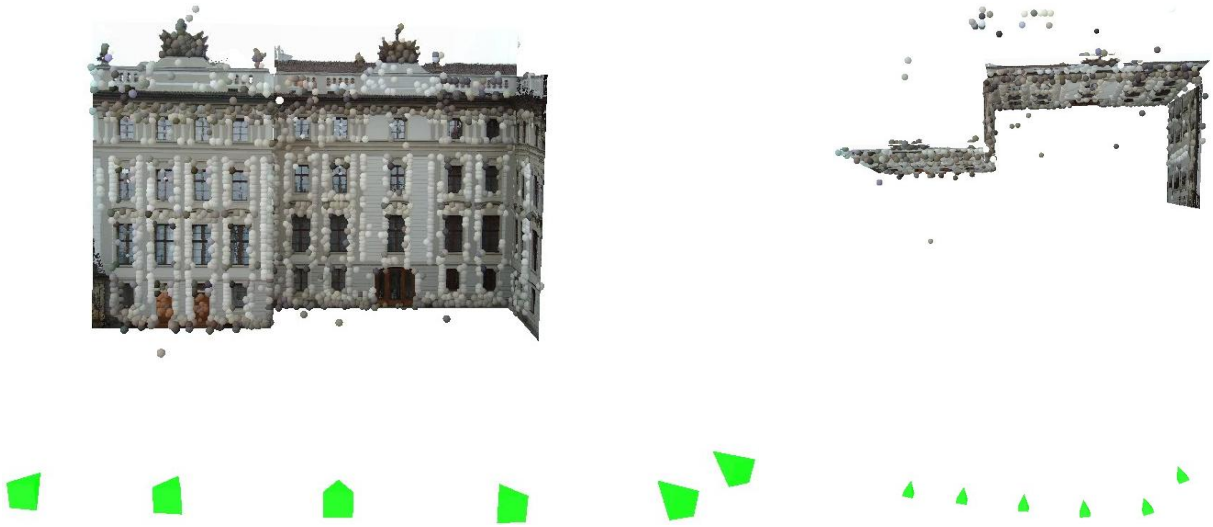

Figure 3. 3D points, colored according to the pixels, facade planes and cameras (green pyramids; cf. Figure 2) derived from the images given in Figure 1 and the 3D model in Figure 2

Figure 4. Facade image derived from the six images given in Figure 1 – left: average; center: median; right: consensus

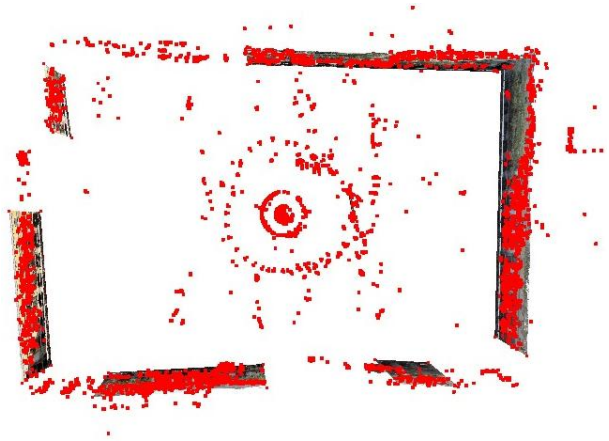

Figure 5. 3D points (red), facade planes, and cameras (medium sized circle around the fountain in the center) derived from nothing but uncalibrated images, ten of them showing the facade on the right hand side given in Figure 8.

the occluding objects, but also because of a not precise enough estimation of the bias of the brightness between the average image and the individual images. The latter stems from the unmodeled occlusions which lead to estimating wrong biases from pixels representing different objects. The latter problem could only be dealt with by robustly recursively estimating biases and occluding objects, which is non-trivial and on our agenda for further research.

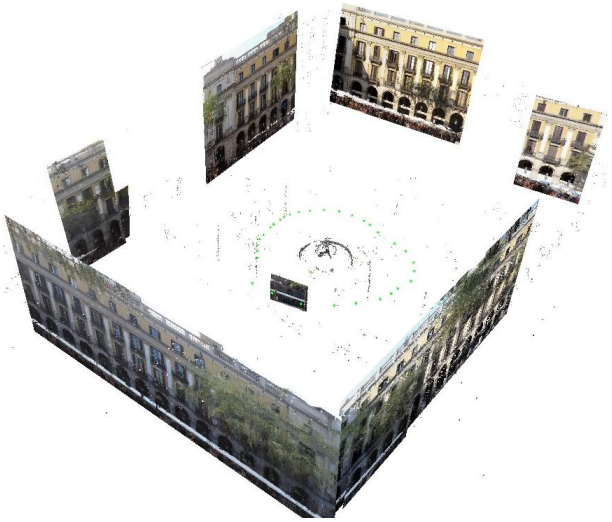

Figure 7. 3D points, colored according to the pixels, facade planes and cameras (green pyramids; cf. Figure 2) from a view above the facade on the right hand side in Figure 5

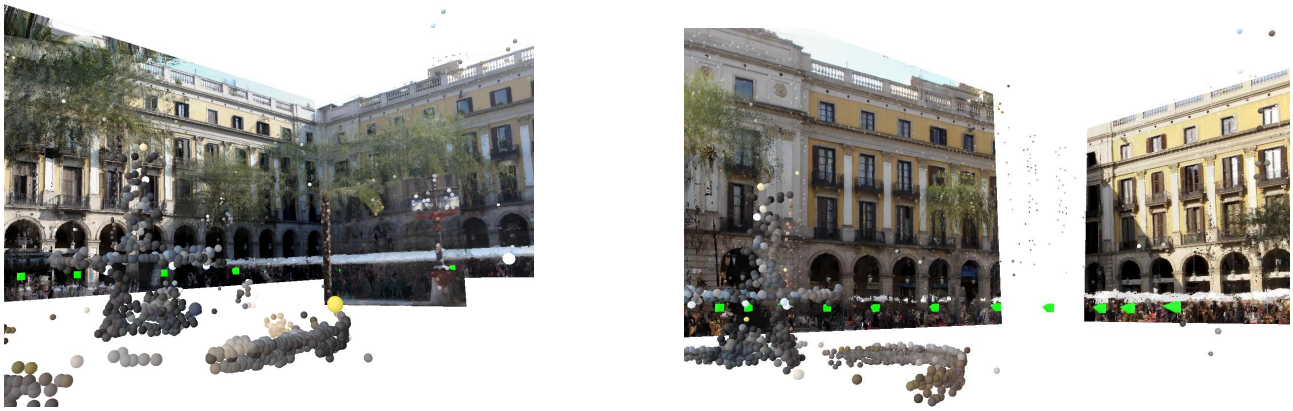

Figure 6. 3D points, colored according to the pixels, facade planes and cameras (green pyramids; cf. Figure 2) from the view of image 11 (left; cf. also Figure 8 upper left image) and 26 (right)

Opposed to the average, both the median and the consensus do much better, even though both are not able to penetrate the vegetation in many instances. If the vegetation is dense, this is not possible at all, but the problem could partly be alleviated by means of more images from different positions. Concerning median and consensus, there are minor, yet characteristic differences. One of the largest can be seen left of the center. The first leaf of the palm tree to the left is mostly eliminated by the consensus, but not by the median, as the former uses redundant information from a larger number of images.

5. CONCLUSIONS

We have shown how combining projective reconstruction with robust techniques and bundle adjustment including covariance information can be used to fully automatically generate textured 3D models of urban scenes from nothing but (possibly uncalibrated) perspective images also for larger numbers of wide-baseline images. These still incomplete 3D models can be the basis for high quality visualizations. Though, at the moment lots of additional manual efforts are needed for a practically satisfying outcome.

One way to proceed is to add detailed geometry by employing semantic information, e.g., by 3D extraction of the windows on the facades (Mayer and Reznik, 2006). After a least-squares fit of the derived planes to all inliers, it will be meaningful to compare the achieved results to ground truth information.

A scientifically interesting path is the SIFT operator (Lowe, 2004) or similar for matching wide-baseline images, as it can deal with orientation and scale differences. The combination with least-squares matching could lead to a broader scope. Though, first tests show that for our limited setup assuming only weak rotation ($\pm 20^\circ$) and scale difference ($\pm 30\%$), we outperform (Lowe, 2004) as we have a more limited search space.

We have experimented with plane sweeping (Baillard and Zisserman, 1999; Werner and Zisserman, 2002), here based on least-squares, to improve the plane parameters derived by RANSAC, but found that for stronger occlusions it is difficult to estimate the bias in brightness. Robust estimation combining, e.g., consensus, with bias determination could be a way to proceed.

Finally, we want to make better use of the information of the planes, e.g., by extending and intersecting planes and checking the newly created planes via homographies, thereby closing gaps. We also plan to employ the intersection lines to improve the determination of the vertical direction, which can be weak for models where mostly walls in one horizontal direction are visible, such as for the Hradschin example.

REFERENCES

Baillard, C. and Zisserman, A., 1999. Automatic Reconstruction of Piecewise Planar Models from Multiple Views. In *Computer Vision and Pattern Recognition*, Volume II, pp. 559–565.

Böhm, J., 2004. Multi Image Fusion for Occlusion-Free Façade Texturing. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume (35) B5, pp. 867–872.

Chum, O., Matas, J., and Kittler, J., 2003. Locally Optimized RANSAC. In *Pattern Recognition – DAGM 2003*, Berlin, Germany, pp. 249–256. Springer-Verlag.

Debevec, P., Taylor, C., and Malik, J., 1996. Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image–Based Approach. Technical Report CSD-96-893, Computer Science Division, University of California at Berkeley, Berkeley, USA.

Dick, A., Torr, P., and Cipolla, R., 2004. Modelling and Interpretation of Architecture from Several Images. *International Journal of Computer Vision* 60(2), 111–134.

Fischler, M. and Bolles, R., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM* 24(6), 381–395.

Förstner, W. and Gülch, E., 1987. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In *ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, Interlaken, Switzerland, pp. 281–305.

Hartley, R. and Zisserman, A., 2003. *Multiple View Geometry in Computer Vision – Second Edition*. Cambridge, UK: Cambridge University Press.

Lhuillier, M. and Quan, L., 2005. A Qasi-Dense Approach to Surface Reconstruction from Uncalibrated Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(3), 418–433.

Lowe, D., 2004. Distintive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2), 91–110.

Mayer, H. and Reznik, S., 2006. MCMC Linked with Implicit Shape Models and Plane Sweeping for 3D Building Facade Interpretation in Image Sequences. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume (36) 3.

Mikhail, E., Bethel, J., and McGlone, J., 2001. *Introduction to Modern Photogrammetry*. New York, USA: John Wiley & Sons, Inc.

Nistér, D., 2004. An Efficient Solution to the Five-Point Relative Pose Problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(6), 756–770.

Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., and Tops, J., 2004. Visual Modeling with a Hand-Held Camera. *International Journal of Computer Vision* 59(3), 207–232.

Torr, P., 1997. An Assessment of Information Criteria for Motion Model Selection. In *Computer Vision and Pattern Recognition*, pp. 47–53.

Wang, X., Totaro, S., Taillandier, F., Hanson, A., and Teller, S., 2002. Recovering Facade Texture and Microstructure from Real-World Images. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume (34) 3A, pp. 381–386.

Werner, T. and Zisserman, A., 2002. New Techniques for Automated Architectural Reconstruction from Photographs. In *Seventh European Conference on Computer Vision*, Volume II, pp. 541–555.

Figure 8. Ten images of Plaza Real in Barcelona from which the facade images given in Figure 9 have been derived

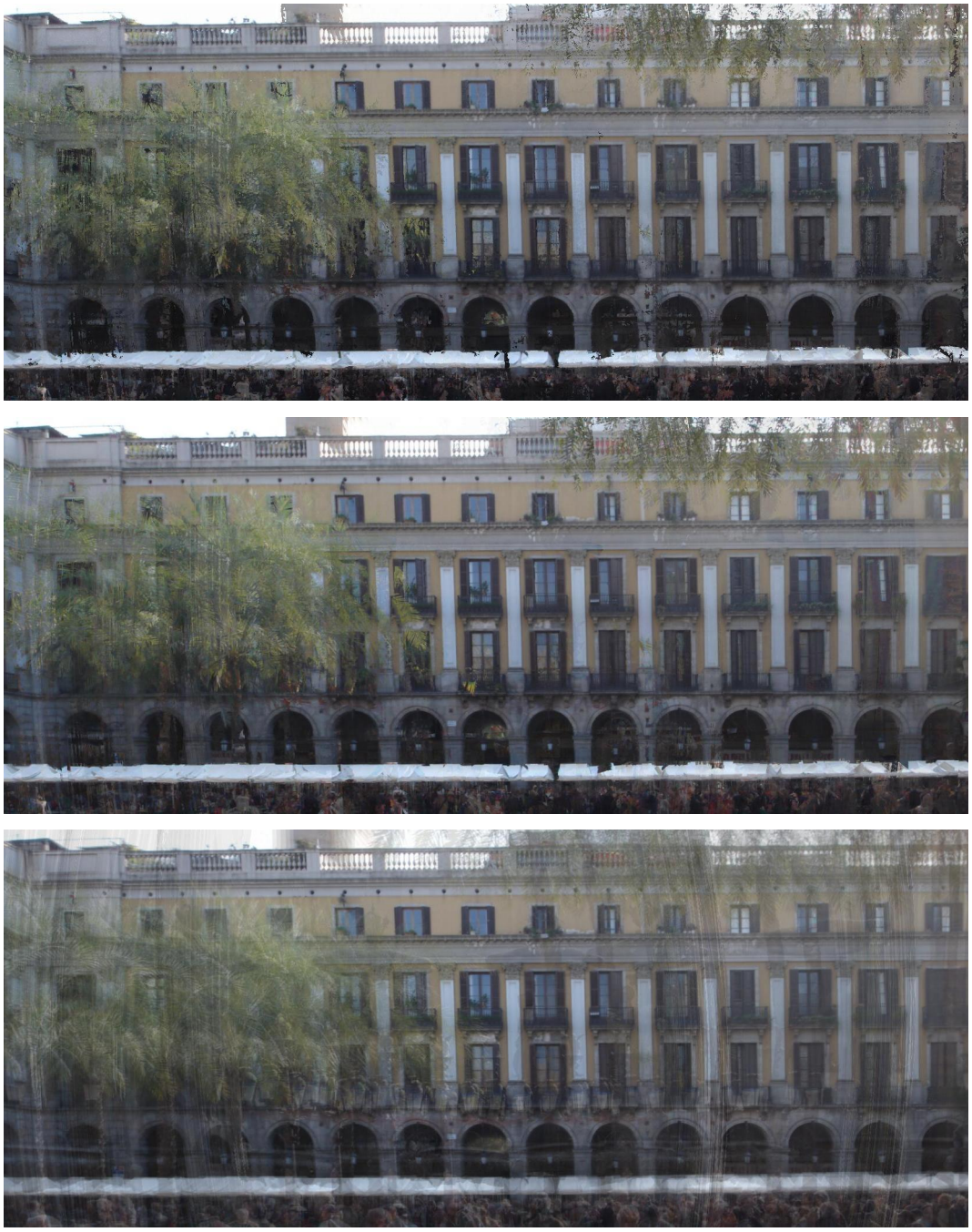

Figure 9. Facade image derived from the ten images given in Figure 8 – bottom: average; center: median; top: consensus