Dense 3D Reconstruction of Symmetric Scenes from a Single Image

Kevin Köser, Christopher Zach, Marc Pollefeys

Computer Vision and Geometry Group, ETH Zürich
Universitätsstrasse 6, 8092 Zürich, Switzerland
{kevin.koeser,chzach,marc.pollefeys}@inf.ethz.ch

Abstract. A system is presented that takes a single image as an input (e.g. showing the interior of St.Peter's Basilica) and automatically detects an arbitrarily oriented symmetry plane in 3D space. Given this symmetry plane a second camera is hallucinated that serves as a virtual second image for dense 3D reconstruction, where the point of view for reconstruction can be chosen on the symmetry plane. This naturally creates a symmetry in the matching costs for dense stereo. Alternatively, we also show how to enforce the 3D symmetry in dense depth estimation for the original image. The two representations are qualitatively compared on several real world images, that also validate our fully automatic approach for dense single image reconstruction.

1 Introduction

Symmetry is a key design principle in man-made structures and it is also frequently present in nature. Quite some effort has been spent to detect or exploit symmetry in computer vision (e.g. [6, 11, 14, 4, 20, 9, 2, 3]). Unlike previous researchers, in this contribution we investigate how 3D symmetry can be exploited to *automatically* obtain a *dense* three-dimensional perception of some scene from a single image, in particular when the scene is symmetric with respect to some virtual symmetry plane. The intuition is that when the observer is not exactly in this symmetry plane, then each object and its symmetric counterpart are seen from a (slightly) different perspective. These two different perspectives onto essentially the same thing can be exploited as in standard two-view stereo to obtain a dense, three-dimensional model (see fig. 1). The key steps are essentially comparable to structure from motion [13], however we run the whole pipeline on a single image, where we assume the intrinsic camera parameters to be known beforehand: Within-image feature matching, robust estimation of autoepipolar geometry and symmetry plane followed by dense depth estimation. Our key contributions are a novel, straight-forward 3D formulation of the single image symmetry scenario which is analog to multi-image structure from motion, a single-texture plane-sweep for a symmetric viewpoint to create a cost volume, and enforcing 3D symmetry in the global optimization by equality constraints in the minimum-surface formulation. The next section will relate this contribution to previous work, before the following sections detail the steps of the approach.

presented at DAGM 2011 (LNCS 6835), see also http://tinyurl.com/depthfromsymmetry

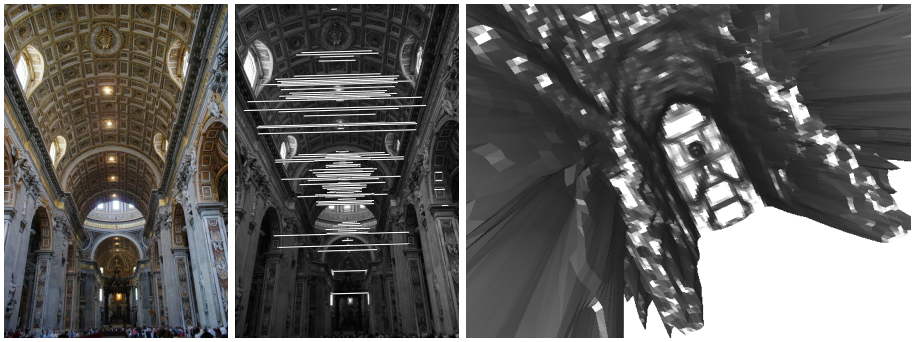


Fig. 1. From left to right: Input image taken at St.Peter's Basilica, detected symmetry (lines connecting features), some rendered oblique view of a very coarse, untextured but shaded model, reconstructed from the single image.

2 Previous work

Previous work related to this contribution can roughly be divided into two categories. The first category describes the general ideas of symmetry exploitation and uses interactive techniques like clicking correspondences or works on restricted scenarios. Gordon [6] seems to be the first to have described the idea of shape from symmetry, later Mitsumoto et al. consider mirrors [11]. Much later, also [4] considers symmetric scene geometry but seems to be unaware of Gordon's work. In terms of dense reconstruction, Shimshoni et al. show interesting results on reconstructing textureless, lambertian objects [14], but they assume a weak perspective camera, horizontal symmetry and a single light source. Their iterative approach starts from a rough estimate of the light source and symmetry plane to optimize normals and scene parameters using shape-from-shading. In other works, geometric relations using planar mirrors have been considered (e.g. [5]) or silhouette-based reconstruction therein [20].

The second category of approaches is concerned about automatic *detection* of symmetry. Here, with the advances of automatic matching, feature-based estimation of planar homologies present in symmetry and repetition (e.g. [15]) became possible, also with non-fronto-parallel planes. In terms of 2D symmetry, Loy and Eklundh [9], observed that SIFT features[8] can either be extracted on mirrored regions or that the SIFT descriptor itself can be rearranged for in-image matching to find mirrored features. Detection was then extended by Cornelius et al. [2, 3] to planar but non-frontal scenes. Wu et al. [16] detected and reconstruct repetitions on rectified planar facades with relief structures. For more details on computational symmetry, we refer to the recent survey paper by Liu et al. [17].

In general, we observe that there is no fully automatic approach to reconstruct a dense, textured 3D model from a single image showing a symmetric scene, which we show in this contribution. Furthermore, we give a novel, clear derivation of the geometry of the symmetric scene which nicely shows the du-

ality of interpreting an image point as being the projection of a reflected point in one camera or the projection of the original point in a "reflected" camera. Finally, we obtain a dense reconstruction that particularly enforces consistent depth for symmetric points. Posing the depth estimation as a labeling problem, we show how to integrate symmetry as equality constraints into a voxel-based (continuous) minimum-cut. The symmetry of the scene allows to compute depth with respect to a central view on the symmetry plane or for the original image, where we compare advantages and drawbacks for both solutions.

3 Symmetric Scene Geometry

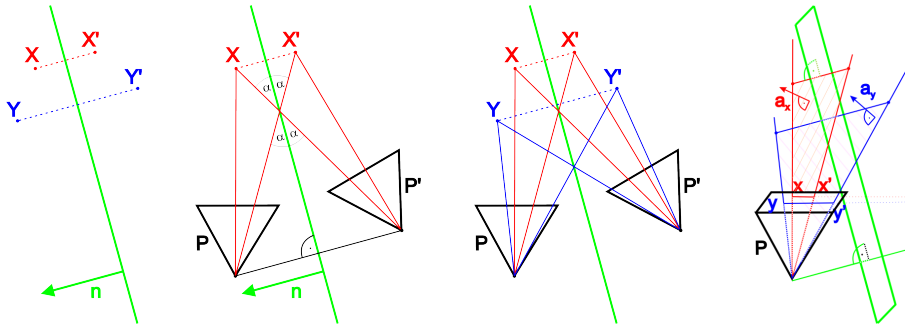


Fig. 2. Left: Reflection of points on a plane with normal $\boldsymbol{n}$ according to equation 2. Center images: dual interpretation of mirrored point and mirrored camera of equation 3. Right: normal constraint of eq.6: the normal $\boldsymbol{n}$ must lie in all backprojection planes of symmetry correspondences.

In this contribution we focus on symmetric scenes, i.e. scenes with a global symmetry plane so that for each point $\boldsymbol{X}$ on one side of the plane there is a corresponding 3D point $\boldsymbol{X}'$ on the opposite side of the plane.

It is easy to see that the image $\boldsymbol{X}'_e$ of a Euclidian 3D point $\boldsymbol{X}_e$ given a mirror plane with normal $\boldsymbol{n}$ through the origin can be obtained by

$$\boldsymbol{X}'_e = \boldsymbol{X}_e - 2(\boldsymbol{n}^\mathsf{T}\boldsymbol{X}_e)\boldsymbol{n} = \left(I_3 - 2\boldsymbol{n}\boldsymbol{n}^\mathsf{T}\right)\boldsymbol{X}_e \tag{1}$$

Now consider that the symmetry plane can have an arbitrary position (not necessarily going through the origin, but by passing it at distance d) and it is expressed in homogeneous coordinates as $\pi = \begin{pmatrix}\boldsymbol{n}^\mathsf{T} & -d\end{pmatrix}$. A reflection by this plane (see figure 2, left image) can then be written linearly in homogeneous coordinates as

$$\mathbf{X}' = \underbrace{\begin{pmatrix} I_3 - 2\boldsymbol{n}\boldsymbol{n}^\mathsf{T} & -2d\boldsymbol{n} \\ \mathbf{0}_3^\mathsf{T} & 1 \end{pmatrix}}_{\mathsf{M}_\pi} \mathbf{X} \tag{2}$$

Here, M_π is a projective transformation that encodes the mirroring.

Consider now a (intrinsically calibrated) camera observing the point $\mathbf{X}$ at image position $\mathbf{x} \simeq \mathsf{P}\mathbf{X}$, where we assume P being the canonic camera at the origin and looking into positive z-direction (cf. to [7]): $\mathsf{P} = (I_3 \;\; \mathbf{0}_3)$. It will observe the mirrored point $\mathbf{X}'$ at $\mathbf{x}' \simeq \mathsf{P}\mathbf{X}'$, which can also be written using M_π as

$$\mathbf{x}' \simeq \overbrace{\mathsf{P} \cdot \underbrace{\mathsf{M}_\pi}}^{\mathsf{P}'} \cdot \underbrace{\mathbf{X}}_{\mathbf{X}'} \tag{3}$$

This equation shows a duality of possible interpretations: M can be absorbed into the projection matrix, defining a new camera that observes an image with $\mathbf{x}$ and $\mathbf{x}'$ swapped, or, M can be absorbed into the point to project the mirrored point (see figure 2, center images). In case we absorb it into the mirrored camera we obtain the 3×4 projection matrix

$$\mathsf{P}' = \mathsf{P}\mathsf{M}_\pi = (I_3 \;\; \mathbf{0}_3) \begin{pmatrix} I_3 - 2\boldsymbol{n}\boldsymbol{n}^\mathsf{T} & -2d\boldsymbol{n} \\ \mathbf{0}_3^\mathsf{T} & 1 \end{pmatrix} = \begin{pmatrix} \underbrace{I_3 - 2\boldsymbol{n}\boldsymbol{n}^\mathsf{T}}_{\mathsf{S}} & -2d\boldsymbol{n} \end{pmatrix} \tag{4}$$

Please note that S is an orthogonal 3×3-matrix with determinant -1 (not a rotation matrix), however P' is still a valid projection matrix. If we compute the essential matrix between P' and P (or between the image and itself) we obtain

$$\mathsf{E} \simeq [\boldsymbol{n}]_\times \left(I_3 - 2\boldsymbol{n}\boldsymbol{n}^\mathsf{T}\right) \simeq [\boldsymbol{n}]_\times \tag{5}$$

which is *autoepipolar* [7].

4 Estimating the Symmetry Plane

Obtaining Correspondences Since the goal is to recover the symmetry plane in three-dimensional scenes, local regions and their symmetric counterparts may look significantly different. In fact, since perspective effects and illumination differences may appear (depending on the distance of the camera to the symmetry plane and depending on illumination and scene normals), this is a wide-baseline matching problem. Similar to previous authors, who were looking for symmetry only in 2D[9] or on planes [3, 2], we exploit the fact that local affine features (cf. to [10] for an overview) locally compensate for perspective effects. Since classical shape + dominant orientation normalization (cf. to [8]) does not allow for general affine transformation but only for those with positive determinant (reflections are not compensated by this), for each feature we explicitly extract also a mirrored descriptor as proposed by [9][1]. Then we find within image correspondences between mirrored and non-mirrored descriptors according to proximity in descriptor space.

[1] If speed is not a concern, just mirroring the whole image along any direction, extracting features and re-assigning descriptors to the coordinates in the original image is sufficient.

Symmetry Plane Normal By definition we know that the line connecting a point $\boldsymbol{X}$ and its symmetric counterpart $\boldsymbol{X}'$ is in direction of the symmetry plane normal (as long as $\boldsymbol{X}$ is not on the symmetry plane and thus identical to $\boldsymbol{X}'$). The plane that is spanned by the camera center and the two viewing rays to the two 3D points contains also this line and consequently the symmetry plane's normal vector must lie in this plane (compare figure 2, right image). Let $(\mathbf{x}, \mathbf{x}')$ be a pair of corresponding (symmetric) features. Then

$$\underbrace{\left([\mathbf{x}']_\times \mathbf{x}\right)^\mathsf{T}}_{\mathbf{a_x}} \boldsymbol{n} = 0 \tag{6}$$

If we use a second correspondence $(\mathbf{y}, \mathbf{y}')$ then the analogue constraint must hold and consequently $\boldsymbol{n}$ has to lie in the null space of the matrix composed of the rows $\mathbf{a}_i$:

$$\begin{pmatrix} \mathbf{a_x} \\ \mathbf{a_y} \end{pmatrix} \boldsymbol{n} = 0 \tag{7}$$

Obviously, $\boldsymbol{n} \simeq \mathbf{a_x} \times \mathbf{a_y}$ fulfills this equation. This is a minimal solution to the 3D symmetry normal from 2 points, which is essentially the same as estimating the epipole for autoepipolar matrices or a vanishing point from images of parallel line segments (cf. to [7]). Please note that this is similar in spirit to [3], however we explicitely write it down for 3D scenes.

Since points on the symmetry plane do not provide constraints, we reject all correspondences with less than 10% image width displacement and apply 2-point RANSAC with the above minimal solver to estimate the symmetry plane normal. Afterwards we apply maximum likelihood estimation for the normal as common also for standard vanishing point estimation approaches [7].

Camera geometry Since we are aiming at dense stereo, some baseline is required to reconstruct the scene geometry. We will now construct a second virtual camera to perform the stereo. Assume for now that the original image has not been taken from exactly inside the symmetry plane. Then, since image-based reconstructions are only up to scale, we can define the baseline of our to cameras to be 2. However, since $\boldsymbol{n}$ and $-\boldsymbol{n}$ are projectively equivalent, there are still two options for the second camera center that need to be resolved, e.g. by checking in which of the configurations the correspondences are in front of the cameras.

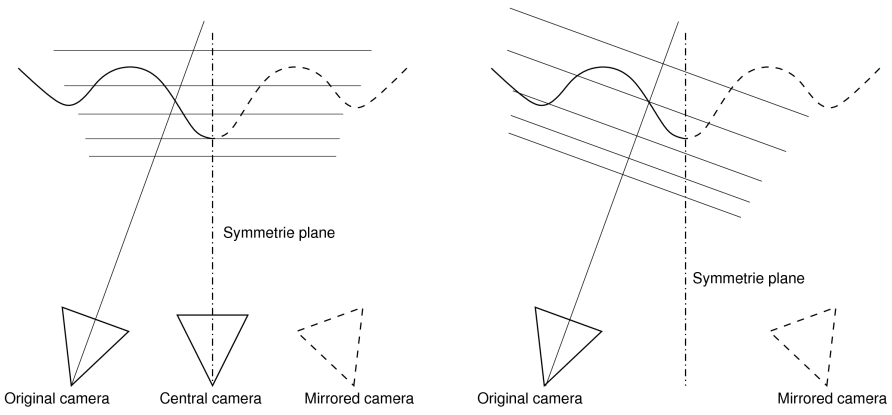
As explained in equation 4, we know that a camera with the projection matrix P' would observe an image with coordinates of $\mathbf{x}$ and $\mathbf{x}'$ swapped. This camera can be converted to a more intuitive right-handed representation by multiplying the projection matrix by -1, subsequent $\mathsf{K}, R, \boldsymbol{C}$-decomposition (this camera is then looking away) and appropriate rotation, e.g. to obtain a rectified standard stereo setup. However, we decided for a more direct approach and use the (left-handed) P' directly in plane-sweep stereo.

We observe that the proposed approach fails in case the camera center is on the symmetry plane, which corresponds to the case of no baseline in standard stereo. Being close to the symmetry plane means only small baseline and potentially only a few measurable disparity steps.

5 Dense Depth Estimation

Dense Stereo and Plane-Sweep Beside using a rectified configuration for dense stereo computation, there are two natural choices for the reference frame used to represent the depth map. Both approaches are based on the plane-sweep methodology. Note that the image to use for the mirrored view is exactly the original image, since symmetric 3D points are treated as the same 3D point by construction. Thus, no additional image has to be synthesized for the mirrored camera.

The first option is to utilize a virtual view between the original and the mirrored camera residing on the mirror plane (see Fig. 3(a)). A fronto-parallel plane sweep approach for stereo with respect to this central view is similar to computational stereo after image rectification, but in this setting both matching images are moving in horizontal direction. This setup has a few advantages, but one major disadvantage. First, by using a symmetric matching score (i.e. symmetric with respect to swapping image patches), and if a symmetric smoothness term is utilized, then the result depth map is naturally symmetric without explicit enforcement. Second, the central view configuration usually minimizes the perspective distortion when using larger aggregation windows for the matching score, since the plane of symmetry is often orthogonal to the surface of man-made objects. Thus, fronto-parallel planes with respect to the central camera tend to be aligned with surface elements leading to better matching scores. The disadvantage of the central reference view configuration is, that there is no fixed reference image unaffected by the current depth hypothesis, and therefore one pixel (or patch) e.g. in the left (original) image may match several pixels/patches in the mirrored view. This leads to noticeable artefacts induced especially by textureless regions (see Fig. 4(b) and (c)).



(a) Central configuration (b) Non-central configuration

Fig. 3. The two setups used for dense depth estimation via plane sweep

The other natural configuration uses the original camera as reference view, and the mirrored camera as matching image (see Fig. 3(b)). The symmetry of the 3D geometry induced by the resulting depth map is lost, and must be explicitly enforced if desired. Since in this configuration the symmetry requirement of the reconstructed object cannot easily be formulated in terms of the resulting depth map, global methods for depth map computation are difficult to extend with symmetry constraints. We utilize the globally optimal stereo method based on finding the minimum-cost surface separating a near plane from a far plane in 3D [1, 12, 18]. In the following section we describe, how 3D symmetry constraints can be incorporated into a class of global stereo methods.

Global Stereo with Symmetry Constraints The basic model for globally optimal stereo is

$$E(u) = \int_{\Omega \times \mathcal{L}} \phi(\nabla u)\, dx\, dl,$$

where $u : \Omega \times \mathcal{L} \to \{0, 1\}$ represents the sublevel function of the desired label assignment $\Lambda : \Omega \to \mathcal{L}$. ϕ is a family of positively 1-homogeneous functions implicitly indexed by grid positions $(x, l) \in \Omega \times \mathcal{L}$. In order to avoid the trivial solution $u \equiv 0$ we fix the boundaries, $u(x, 0) = 0$ and $u(x, L) = 1$.

Each grid position (x, l) (i.e. a camera ray with an associated depth label) corresponds to a point X in 3D space. Thus, u can be interpreted as 3D occupancy function whether a voxel corresponding to (x, l) is filled ($u(x, l) = 1$) or empty ($u(x, l) = 0$). Knowing that the object to be modeled is symmetric with respect to a mirror plane $\boldsymbol{n}^\mathsf{T}\boldsymbol{X} = 1$, implies that both 3D locations $\boldsymbol{X}$ and its reflection $\boldsymbol{X}' = (\boldsymbol{I} - 2\boldsymbol{n}\boldsymbol{n}^\mathsf{T})\boldsymbol{X} + 2\boldsymbol{n}$ are either occupied or empty, i.e. have the same state. The constraints can be translated to a set of equality constraints for corresponding locations in the domain $\Omega \times \mathcal{L}$, $u(x, l) = u(x', l')$ for $\big((x, l), (x', l')\big) \in C$. Here C is a set of corresponding locations within the viewing frustum.

Overall the depth labeling task can be written as (after relaxing the binary constraint $u(x) \in \{0, 1\}$ to $u(x) \in [0, 1]$)

$$E(u) = \int_{\Omega \times \mathcal{L}} \phi(\nabla u)\, dx\, dl,$$

subject to $u(x, l) \in [0, 1]$, $u(x, 0) = 0$, $u(x, L) = 1$, and $u(x, l) = u(x', l')$ for $\big((x, l), (x', l')\big) \in C$. This is a non-smooth convex problem. On a discrete grid and with ϕ being a weighted L^1 norm this can be solved with a graph cut method [1]. The additional equality constraints can be enforced by infinite links between the respective nodes in the graph (and therefore both sites have to be on the same side of the cut). Using similar arguments as in [19] it can be shown that u attains essentially binary values also in the case of general positively 1-homogeneous functions ϕ.

Implementation Details We utilize the L^1 difference between 5×5 pixel patches of Sobel-filtered images as our image matching score. Consequently, pixel brightness differences due to shading effects are largely addressed by this choice of the similarity function. We increase the robustness of the matching score by truncating its value to a maximum of 5 (with respect to normalized pixel intensities in $[0, 1]$). This is very helpful to limit the influence of non-symmetric high-frequency textures on the overall result. Global optimization to obtain a smooth depth map is based on a primal-dual gradient method. Spatial smoothness is enforced by utilizing the isotropic total variation. The final depth value (potentially at subpixel accuracy) is extracted from u as the 0.5-isolevel.

6 Experiments

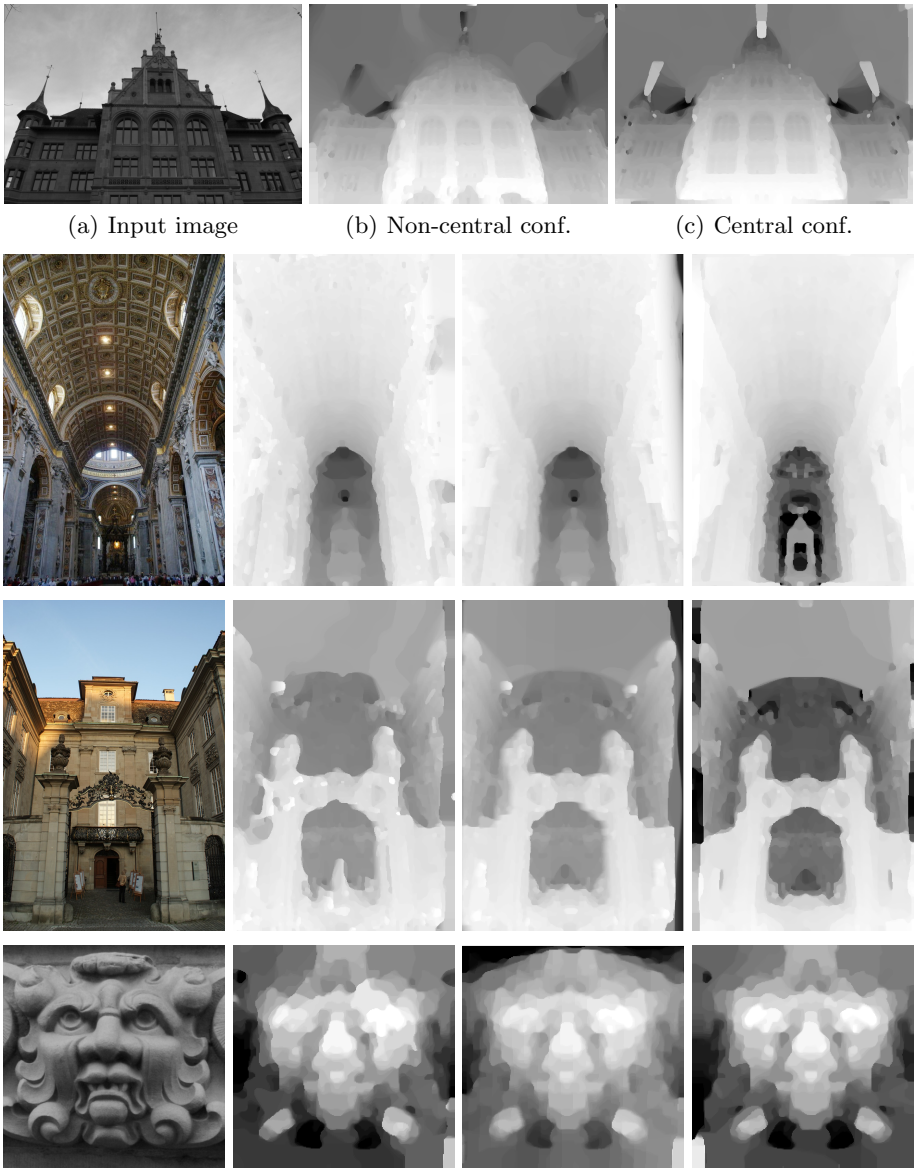
We evaluated our approach on a set of real images from a range of several scenarios (see Fig. 4): facades, indoor environments with a large depth range, depth discontinuities and occlusions, and finally rather textureless and only approximately symmetric objects. In order to cope with inaccuracies of the estimated symmetry plane normal, and to be robust with respect to texture asymmetries at small scales, we downsized the images to quarter resolution (of originally 3-6 MegaPixel). The plane sweep method evaluates 120 depth values, and the weight parameter for the data fidelity term is set to 5.

Since a quantitative evaluation is difficult due to missing ground truth, we qualitatively compare the different approaches. First it can be observed that the global but unconstrained solution does not produce symmetric 3D scenes, whereas the other approaches do. Qualitatively the depth maps returned by the different dense stereo methods are similar, although they differ in details. While the central approach seems to be attractive because of its intrinsic 2D symmetry of the depth map, we noticed that it can introduce undesired artefacts: in the central configuration a single pixel of the original view can be consistent with different depth hypotheses and thus be assigned to multiple depths. Objects with small depth variations and concave environments are clearly most suitable for symmetry-based single view reconstruction, due to the absense of strong occlusions (Fig. 4, first two rows).

7 Conclusion

After anaylzing the underlying 3D geometry, we have presented a novel automatic approach to densely reconstruct a symmetric scene from a single image. In particular we suggested and compared different representations of the 3D scene (depth with respect to a virtual central view or with respect to the original camera), and enforced the reconstructed scene to be symmetric by equality constraints between corresponding 3D locations in a minimal surface formulation.

Future work might exploit multiple local symmetries, and could also investigate in detecting the support of the detected symmetry in the image, i.e. separate symmetric and non-symmetric scene elements.



(a) Input image (b) Non-central conf. (c) Central conf.

(d) Input image (e) Unconstrained (f) Symm. constr. (g) Central conf.

Fig. 4. Results for our datasets. First row: while the central configuration naturally leads to symmetric depth maps, some artefacts induced by textureless regions are visible (see the apexes of the towers in (b) and (c)). Bottom rows: input images (d) and depth maps obtained for the non-central configuration without explicit symmetry constraints (e), with symmetry constraints (f), and for the central configuration (g).

References

1. Boykov, Y., Veksler, O., Zabih, R.: Markov random fields with efficient approximations. In: Proc. CVPR. pp. 648–655 (1998)
2. Cornelius, H., Loy, G.: Detecting bilateral symmetry in perspective. In: Proceedings of the 2006 CVPR-Workshop on POCV. p. 191ff. (2006)
3. Cornelius, H., Perd'och, M., Matas, J., Loy, G.: Efficient symmetry detection using local affine frames. In: SCIA. pp. 152–161 (2007)
4. Francois, A., Medioni, G., Waupotitsch, R.: Reconstructing mirror symmetric scenes from a single view using 2-view stereo geometry. In: Pattern Recognition, 2002. Proceedings. 16th International Conference on. vol. 4, pp. 12 – 16 vol.4 (2002)
5. Fujiyama, S., Sakaue, F., Sato, J.: Multiple view geometries for mirrors and cameras. In: Proceedings of ICPR. pp. 45 –48 (2010)
6. Gordon, G.: Shape from symmetry. In: Proc. of SPIE, Intelligent Robots and Computer Vision VIII: Algorithms and Techniques. vol. 1192 (1989)
7. Hartley, R., Zissermann, A.: Multiple View Geometry in Computer Vision. Cambridge university press, second edn. (2004)
8. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
9. Loy, G., Eklundh, J.O.: Detecting symmetry and symmetric constellations of features. In: ECCV. pp. 508–521 (2006)
10. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A Comparison of Affine Region Detectors. International Journal of Computer Vision 65(1-2), 43–72 (2005)
11. Mitsumoto, H., Tamura, S., Okazaki, K., Kajimi, N., Fukui, Y.: 3-d reconstruction using mirror images based on a plane symmetry recovering method. Pattern Analysis and Machine Intelligence, IEEE Transact. on 14(9), 941 –946 (1992)
12. Pock, T., Schoenemann, T., Graber, G., Cremers, D., Bischof, H.: A convex formulation of continuous multi-label problems. In: Proc. ECCV (2008)
13. Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R.: Visual modeling with a hand-held camera. IJCV 59, 207–232 (2004)
14. Shimshoni, I., Moses, Y., Lindenbaum, M.: Shape reconstruction of 3d bilaterally symmetric surfaces. In: Int. Conf. Image Analysis and Processing. pp. 76 –81 (1999)
15. Tuytelaars, T., Turina, A., Gool, L.V.: Non-Combinatorial Detection of Regular Repetitions under Perspective Skew. IEEE Trans. PAMI 25(4), 418–432 (2003)
16. Wu, C., Frahm, J.M., Pollefeys, M.: Repetition-based dense single-view reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)
17. Yanxi Liu, Hagit Hel-Or, C.S.K., Gool, L.V.: Computational symmetry in computer vision and computer graphics. Foundations and Trends in Computer Graphics and Vision 5(1-2), 1–195 (2010)
18. Zach, C., Niethammer, M., Frahm, J.M.: Continuous maximal flows and Wulff shapes: Application to MRFs. In: Proc. CVPR. pp. 1911–1918 (2009)
19. Zach, C., Shan, L., Niethammer, M.: Globally optimal Finsler active contours. In: Pattern Recognition. LNCS, vol. 5748, pp. 552–561 (2009)
20. Zhong, H., Sze, W., Hung, Y.: Reconstruction from plane mirror reflection. In: Proceedings of ICPR 2006. vol. 1, pp. 715 –718 (2006)