Presented at 2nd Workshop on Statistical Methods in Video Processing, SMVP 2004, Prague, May 2004.
This is a pre-print of the paper that will appear in Springer LNCS proceedings.

Towards Complete Free-Form Reconstruction of Complex 3D Scenes from an Unordered Set of Uncalibrated Images

H. Cornelius[1], R. Šára[2], D. Martinec[2], T. Pajdla[2], O. Chum[2], and J. Matas[2]

[1] Royal Institute of Technology (KTH)
Department of Numerical Analysis and Computing Science
100 44 Stockholm, Sweden
hugoc@nada.kth.se, http://www.nada.kth.se
[2] Center for Machine Perception, Czech Technical University
166 27 Prague, Czech Republic
{sara,martid1,pajdla,chum,matas}@cmp.felk.cvut.cz, http://cmp.felk.cvut.cz

Abstract. This paper describes a method for accurate dense reconstruction of a complex scene from a small set of high-resolution unorganized still images taken by a hand-held digital camera. A fully automatic data processing pipeline is proposed. Highly discriminative features are first detected in all images. Correspondences are then found in all image pairs by wide-baseline stereo matching and used in a scene structure and camera reconstruction step that can cope with occlusion and outliers. Image pairs suitable for dense matching are automatically selected, rectified and used in dense binocular matching. The dense point cloud obtained as the union of all pairwise reconstructions is fused by local approximation using oriented geometric primitives. For texturing, every primitive is mapped on the image with the best resolution.
The global structure reconstruction in the first step allows us to work with an unorganized set of images and to avoid error accumulation. By using object-centered geometric primitives we are able to preserve the flexibility of the method to describe complex free-form structures, preserve the possibility to build the dense model in an incremental way, and to retain the possibility to refine the cameras and the dense model by bundle adjustment. Results are demonstrated on partial models of a circular church and a Henri de Miller's sculpture. We observed spatial resolution in the range of centimeters on objects of about 20 m in size.

1 Introduction

Building geometric representation of a complex scene from a set of views is one of the classical Computer Vision problems. The task is to obtain a model that (1) can either be used to generate a novel view for a moving observer or (2) contains explicit representation of the structure (3D topology and geometry) of the scene. The focus of this paper is on the latter. We present a method that obtains the 3D model from a small unordered set of uncalibrated images. This means that the

camera order, positions and (most of) their intrinsic parameters are not known. Hence, a method that is capable of *joint* estimation of the cameras and the scene structure must be used. This is usually done in two independent steps: a pre-calibration, which recovers the cameras from a sparse set of features, followed by scene structure estimation and densification. To increase the accuracy of both scene structure and the cameras, this can be followed by bundle adjustment.

In this paper we basically follow the cascade: (1) wide-baseline matching, (2) camera and 3D structure reconstruction, (3) dense matching, (4) 3D model reconstruction. By splitting the procedure to smaller, explicit blocks, we hope for achieving good performance by solving a well defined task at every step. Every such task can then use a different, optimal, prior model of the scene. The first stage uses discriminative image regions and a strong local planarity model to establish initial matches that are used in the second stage, where a consistent set of cameras are searched for under occlusion and outliers. Pinhole camera model with radial distortion is used. The subsequent matching then focuses on density and accuracy while minimizing false positives under missing texture and inaccurate epipolar geometry. The last step first approximates the point cloud density as a mixture of local kernels which helps maintain the efficiency of the subsequent processing. Since every stage makes only a small step in image interpretation, in our future work we will be able to close a feedback loop from almost every stage and refine the camera parameters and the scene structure before we commit to final interpretation of scene structure (like a triangulated model of certain topology).

Our system, described in detail in this paper, differs from other work by:

1. A new method for the reconstruction of the projective structure and the cameras using a *robust* and *global* optimization procedure that does not require the input images to form a sequence (a known-order image set). This means that reconstruction errors do not propagate.
2. Using an *object-centered* model of the *local* geometry of the scene, which gives the possibility (1) to process and model images of complex structures, (2) to grow the model in an incremental way while preserving the accuracy, (3) to insert high-resolution partial reconstructions into the model, (4) to perform an efficient iterative refinement (bundle adjustment) in the final accuracy-increasing step, (5) to impose local spatial coherence, and (6) to effectively *compress* the dense point cloud while preserving its ability to model complex free-form structures.

2 Method

In this section we describe the data processing pipeline. In brief the method works as follows. The input is a number of photographs of the object to be reconstructed. Sparse correspondences are searched for across all pairs of views. A consistent system of cameras is estimated and the cameras are auto-calibrated. Image pairs suitable for dense matching are selected and rectified and used for dense matching. 3D points are reconstructed from each image pair and the union

of the partial reconstructions form a dense point cloud. Local geometric models called fish-scales are fit to the point cloud and each fish-scale is texture-mapped from the image with the best resolution and view of that fish-scale.

Input The input is a number of photographs of the object to be reconstructed, taken with a hand-held compact digital camera. For the method to work well, there should be image pairs taken with both wide and relatively narrow baseline among the images. The wide baseline pairs support the numerical stability of camera auto-calibration. If available, the narrow baseline pairs are more suitable for dense matching because stronger prior models (like ordering) can be used.

Our method makes it possible to use photographs of different resolution. To reconstruct the overall 3D structure of the object, overview images are used. Higher resolution images can be used to blend in parts of the object with fine geometric details or with texture too poor to suffice for reliable matching at the lower resolution. Examples of input data obtained under this scheme are shown in Figs. 1 and 2.

Fig. 1. Input images for the Head Scene

Region matching The first step is to find sparse correspondences across all images. This is done by matching maximally stable extremal regions (MSERs) [1] in all possible pairs of views. The epipolar geometry for each image pair is estimated using LO-RANSAC [2]. Taking only the matches satisfying the epipolar constraint, we get for every image pair a set of sparse correspondences with relatively few outliers with respect to the true scene structure. The ability of the method to handle large changes in scale and brightness is essential, since we are necessarily dealing with wide baseline photographs of varying resolution.

Suppose the object has two or more parts looking the same. To reduce the risk that too many matches between similar regions on different parts of the object would result in a wrong reconstruction of the cameras, it might seem necessary to forbid matching between images seeing these different parts. However, it is possible to phase-out most of such image pairs automatically, as described below.

An example of a set of detected MSERs in a wide baseline pair is shown in Fig. 3.

Fig. 2. Input images for the St. Martin scene. Note that there are narrow-baseline overview and close-up pairs that are mutually separated by wide baselines

Fig. 3. Maximally stable extremal regions (circles) detected in two wide-baseline pairs

Estimation of a consistent system of cameras Assuming full perspective camera model, projection of each point $\mathbf{X}_p$ visible in camera $\mathtt{P}^i$ can be written in homogeneous representation as $\lambda_p^i \mathbf{x}_p^i = \mathtt{P}^i \mathbf{X}_p$ where λ_p^i is a non-zero scale called *projective depth*. Projections of all points into all images can be gathered into one large matrix equation $\mathtt{M} = \mathtt{PX}$ where $\mathtt{M}$ is so called *rescaled measurement matrix* which contains images of all points rescaled by projective depths, $3m \times 4$ matrix $\mathtt{P}$ contains m camera matrices stacked on top of each other, and a $4 \times n$ matrix $\mathtt{X}$ contains n 3D points. $\mathtt{M}$ has one column per 3D point and three rows per camera. If some point is not visible in some camera, the corresponding entries in matrix $\mathtt{M}$ are unknown. Projective structure, $\mathtt{X}$, and motion, $\mathtt{P}$, can be found by factorizing this large matrix. We use a modification of method [3] which is able to deal with both occlusions (missing entries) and outliers. This is necessary since we want to be able to do full reconstructions of objects of any shape, and since there may always be outliers among the matched points. Note that the method does not put any restrictions on image order.

All inliners with respect to the epipolar geometry, i.e. the pair-wise matches obtained from the region-matching step, can be placed into the $\mathtt{M}$-matrix. In

the original method [3], the conflicting matches are simply ignored and outliers removed in a subsequent stage using trifocal tensors. However, this turned out not to work when there were many image pairs with no mutual overlap. There will always be some matches between these image pairs and (incorrect) epipolar geometries (EGs) will be estimated with usually just a few matches satisfying them. Still, the number of matches may be higher than in some other image pair with a correct EG but with only a few matches due to small image overlap. Therefore, discarding pairs with the number of matches falling below some threshold does not work. A simple greedy algorithm can overcome this difficulty: First, matches from the image pair with the most inliers with respect to the EG are loaded into the `M`-matrix. Next, matches from the pair with the second largest number of inliers are loaded etc. This guarantees that more reliable matches are used first. Each match is checked against the already loaded EGs and if it satisfies them, it is merged into the `M`-matrix. It turns out that many outliers and only a few inliers are discarded this way.

Camera auto-calibration The reconstruction obtained in the last step is projective. To upgrade the projective reconstruction to a metric one, the cameras are auto-calibrated using the image of the absolute dual quadric [4]. The constraints used for the calibration are square pixels and zero skew. When more information about the cameras is available, more constraints can be used. To improve the quality of the solution, the calibration is followed by bundle adjustment including a radial distortion model.

Radial distortion correction Since real cameras deviate from the linear pinhole model, the images have to be corrected for radial distortion. This is done by unwarping the images using the radial distortion model estimated in the bundle adjustment step described above. We use the division model $r_p = \frac{r}{1+\lambda r^2}$, where r_p is the perfect (undistorted) radius (measured from the distortion center $\mathbf{x}_0$), r the distorted radius and λ the radial distortion parameter. See [5] for the properties of this model. The reasons for using this model are that it is simple, performs well and that its inverse has a simple closed form except for at $r_p = 0$.

Image pair rectification Rectification is necessary for an efficient dense matching procedure. After radial distortion rectification described above, the image pairs that will be used for dense matching are rectified by applying homographies mapping the epipoles to infinity on the horizontal axis [4]. This approach does not work if the epipoles are inside the images or too close to the image borders. Therefore image pairs, for which this is true are excluded from dense matching in an automatic close-pair selection step just before the rectification. The area around the epipoles would not provide 3D reconstructions of good geometric accuracy anyway and linearly rectified image pairs are more suitable for sub-pixel disparity estimation because of the simplicity of the underlying model which makes the algorithm faster and numerically better posed.

Dense matching Dense matching is performed as a disparity search along epipolar lines using Confidently Stable Matching (CSM) [6]. This algorithm assumes that the ordering constraint holds. If it does not, the corresponding part

of the scene is rejected in the respective image pair. The CSM was used because is has a very low mismatch rate [7] and is fast. The single important parameter to CSM, for which a default value cannot be used, is the disparity search range. We set this parameter to the range of the known disparities of the sparse MSER matches plus a fixed margin. A typical search range for the overview images in the St. Martin scene (the church) is ± 100 pixels. The algorithm has two more parameters: α, used to reject insufficient signal-to-noise ratio image data and β, used for repetitive pattern rejection (see [6]). By construction of the CSM algorithm none of these is critical nor scene-dependent. These parameters are both set to default values.

The output from the matching algorithm is one disparity map per image pair admitted for dense matching (see Fig. 4). By least squares estimation using an affine distortion model the disparity maps are upgraded to sub-pixel resolution [8].

Fig. 4. The disparity map for the first two images in the second row in Fig. 2 and Point clouds for the St. Martin scene (a front and a top view). Only 2% of all points are shown

Point cloud reconstruction and local aggregation to fish-scales From the disparity maps the corresponding 3D points are reconstructed. The union of the points from all disparity maps forms a dense point cloud (see Fig 4).

An efficient way of representing distributions of points is to use fish-scales [9]. Fish-scales are local covariance ellipsoids that are fit to the points. They can be visualized as small round discs (see the results in Figs. 5, 6 and 7). A collection of fish-scales approximate the spatial density function of the measurement in 3D space.

The most important parameter of the fish-scale fitting is the fish-scale size. A too small value results in a noisy and sparse model and a too large value does not model fine structures well. The appropriate fish-scale size is found by sampling the density of the point cloud in the neighborhood of a number of points and

by grounding the fish-scale size on the median point density. Here we use one fish-scale size for modeling the overall structure of the object, and a smaller size for modeling the details reconstructed from the high-resolution images. One point cloud from the low and one point cloud from the high resolution images are reconstructed and fish-scales are fit to the two point clouds independently. The large fish-scales that are close to a small fish-scale are rejected, and the two results are then fused by their union.

Texturing Texture can easily be mapped on the fish-scales. However, we first have to decide from which view to get the texture for a certain fish-scale. This is done by counting the number of points reconstructed from each image pair within a certain Mahalanobis distance from a fish-scale. The number of points per view (one view could be used in several image pairs) are counted and by taking the view with the highest number of points we get the image with the best resolution and best view of the fish-scale. The method for choosing the best view for texturing takes one parameter, the distance within which to count the points.

3 Experiments and Discussion

We have applied our 3D reconstruction method to three different scenes: the St. Martin rotunda at Vyšehrad in Prague, the sculpture "l'Ecoute" (Listening) by Henri de Miller in Paris, and the Valbonne church near Nice. Image sizes were about 1000×1400 pixels for the rotunda and the sculpture and 512×768 for the church. The used input images can be seen in Figures 2, 1, and 7.

The images of the St. Martin rotunda were specially acquired for the purpose of 3D reconstruction. Overall images capturing the whole or major parts of the building were taken in such a way that views from adjacent positions would have a reasonable overlap. From each shooting position a narrow baseline image pair was taken. The baselines were about 1 to 1.5 meters. From some of the positions zoomed-in images of areas with fine geometric structures or poor texture were taken. For the experiment presented in this article only a subset of the images taken were used (see Fig. 2). The Head images are not optimal, they form a simple semicircular sequence (see Fig. 1). The Valbonne images form two semicircular trajectories (see Fig.7).

For all scenes, the whole procedure was performed fully automatically and with the same parameters. The only prior knowledge used was which focal lengths were the same: For the Head scene we knew that the focal length and principal point were approximately the same for all images. For the St. Martin scene the focal length and principal point were the same within the narrow baseline pairs. This information was used when auto-calibrating the cameras. However, if this knowledge was not used, very similar results were obtained. No knowledge of internal camera parameters was available for the Valbonne scene.

The region matching could be done for all image pairs and no pairs had to be manually forbidden. Some matches were always found, although some image

pairs had no overlap. However, these matches were quite few and our method for finding the cameras was able to deal with them.

The narrow-baseline pair selection was quite simple: First, all pairs not suitable for rectification are forbidden. Next, the image pair with the highest number of inliers from the MSER matching is chosen. After that the second best pair is added and so on. When choosing a new pair, we require that at least one of the images in the pair has not been used before. This way every image is matched to the best image possible. Although non-optimal, the method gave the desired result for the St. Martin scene plus one extra pair (7–9) and also a good result (1–2, 2–3, 3–4, 5–6, 7–8, 9–10) when applied to the Head scene images.

The dense matching was carried out as described above. At least 30 points per volume of fixed size was required to make a fish-scale. Results for the three scenes are shown in Figs. 5, 6 and 7.

Fig. 5. The fish-scale model (textured and untextured) for the Head scene

Discussion In all scenes, the models are smooth, and curvature is well captured. The reconstructions are highly accurate. For example, the pilaster on the apse of the rotunda is clearly visible in the reconstruction although it is only around 30 cm wide and less than 10 cm on the side and reconstructed from photographs taken approximately 20 meters away with a one meter baseline. Note also the thin structures like the ball at the cupola of the church. No jumps are visible on the boundaries between parts of the object reconstructed from different image pairs. The low-resolution fish-scales are aggregated from one single point cloud consisting of the points from all low resolution image pairs. Hence, any possible

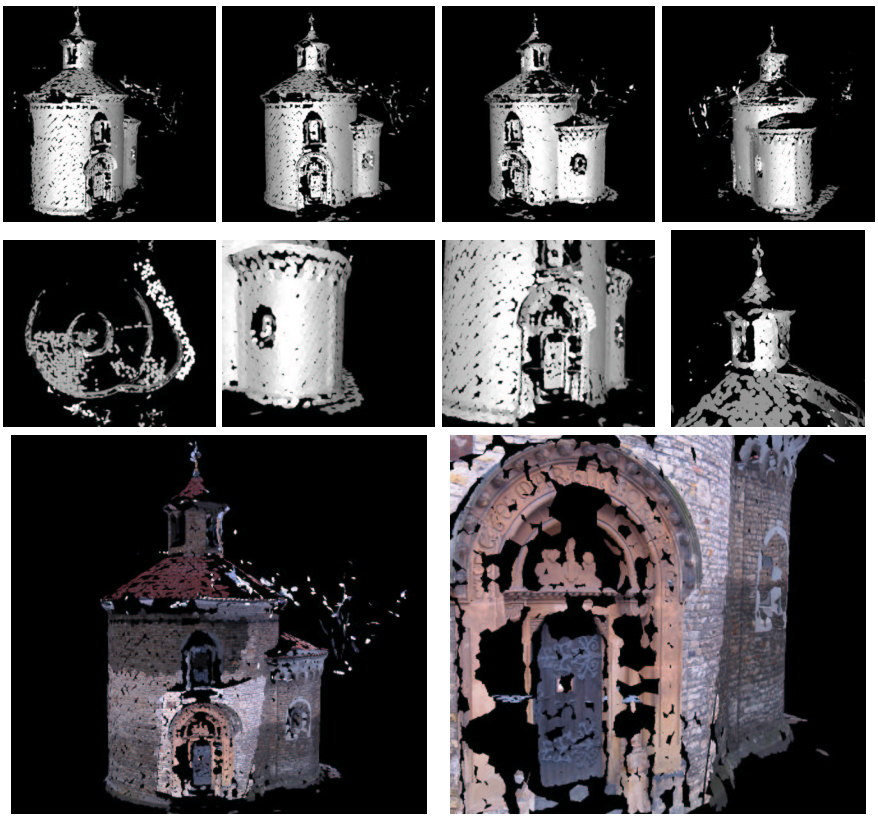

Fig. 6. The fish-scale model for the St. Martin scene. Note the ball at the cupola and parts of the trees around the church. The first image in the second row is the top view. Texture is not radiometrically corrected to demonstrate which views contributed

gaps along image borders would be smoothed-out. The blended-in details, on the other hand, are aggregated independently, from a different (denser) point cloud, still no gaps are visible on borders between the two fish-scale sets, see the region around the door in Fig. 6.

The current version of fish-scale rendering has problems to capture thin and branching structures accurately, the rendering makes them more flat than in the actual model. This is visible on the ball at the cupola and on the branches of the surrounding trees.

In our approach we recover only spatial features that have *strong support in data*. Computational resources are not wasted to densely explain all—even weak-texture—data at once as in global optimization methods [10–12]. Fish-scales reconstructed from low-information patches would not contribute to the accuracy while increasing the demand for greater computational resources.

The fish-scales are considered as an intermediate 3D model. They capture local surface, including its orientation very efficiently. This makes them suitable

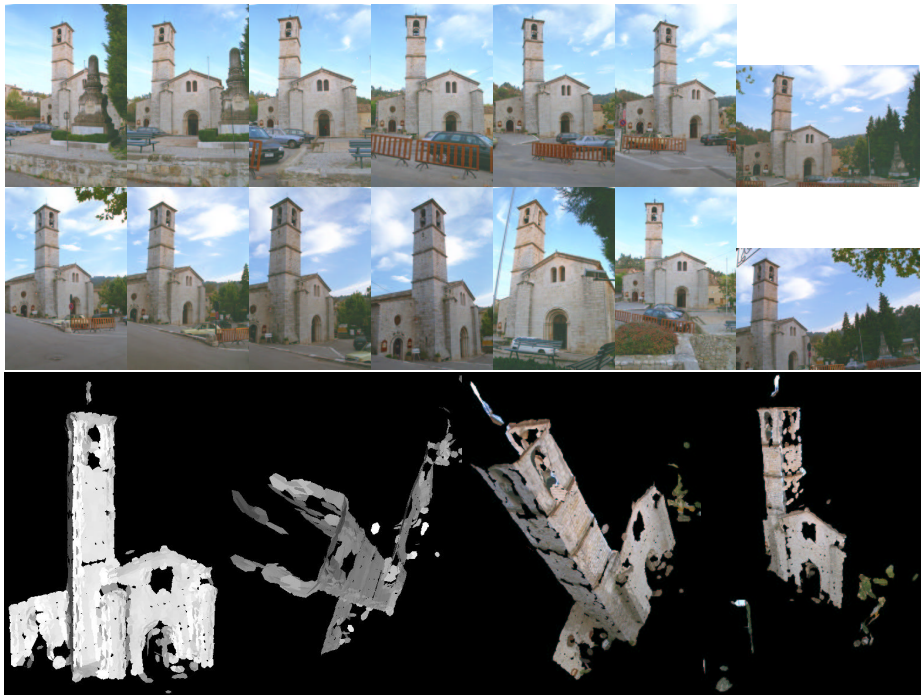

Fig. 7. Input images and the fish-scale model for the Valbonne scene. Note that the walls of the tower are mutually perpendicular

for further camera parameter improvement, jointly with the 3D structure estimation. The result of such an iterative procedure could then be interpreted as a triangulated surface as, for instance, in [13, 14, 9] or rendered directly as in [15]. Alternatively, the fish-scale model could be used as an initialization to a global optimization procedure that relies on high-accuracy camera calibration.

Our future work will include full-complexity fusion of partial reconstructions that requires selection of the best models on the partial model overlap. High-accuracy fish-scales should have precedence over low-accuracy fish-scales obtained from coarser-resolution images. Another part to be finished is bundle adjustment of the fish-scale model. In a final step we would like to densify or 'extend' the set of fish-scales by generating hypotheses around boundaries of the existing fish-scale sets and validate/refine them in the images, as in [16]. This possibility has been studied in [17]. We would also like to improve the close pair selection algorithm. In a large dataset it takes a very long time to run the MSER matching between all pairs of views. We study the possibility to avoid running the matching between all pairs. For example an approach similar to the one presented in [18] could be used.

4 Summary and Conclusions

In this work we have shown that given a wide baseline stereo matching algorithm, an occlusion-robust algorithm for estimating a consistent system of cameras from pair-wise point correspondences, and a dense stereo-matching algorithm, it is possible to obtain automatic high-resolution metric 3D reconstructions of objects of complex shape from a set of photographs. The strong requirement on the object to be reconstructed is that it must have sufficient texture, since this is required for the dense matching algorithm and for the geometric accuracy of the result. The requirements on the images are that among the photographs there have to be some image pairs suitable for dense matching and some wide baseline photographs to support the numerical stability of the camera calibration. By using image pairs of different resolutions it is possible to reconstruct the overall shape of the object from images with one resolution and to use higher resolution photographs for important details or poorly textured areas.

For the success of the data processing pipeline we find the following critical:

1. To obtain valid camera reconstructions there must be enough discriminatory regions (like the MSER) for the initial matching. This success is scene-dependent. Detecting sigle-type discriminatory regions need not suffice if the scenes are unconstrained in appearance.
2. The camera and scene reconstruction module has to cope with severe occlusion and moderate fraction of outliers in data.
3. Dense matching must produce few mismatches.
4. Surface reconstruction has to cope with complex structures, holes and missing data, and with a small to moderate fraction of outliers.

All points except for the first are satisfied in the method described here.

For the accuracy of the result the following is not critical but important: (1) Subpixel disparity estimation. (2) Radial distortion modeling. (3) Accurate epipolar geometry estimate before dense matching.

In this work we were surprised by the following: (1) The naroweness of the baseline for dense matching is not detrimental to the accuracy of the final model. (2) Higher resolution model parts did not require any additional effort to be blended in seamlessly. (3) It was possible to reconstruct thin structures consistently and with good accuracy, even from many uncalibrated views and even at the extremities of the scene like the ball at the cupola of the church. (4) Disparity maps need not be of full density to recover the fish-scale model because holes in one image pair are usually covered from another pair. (5) The method does not break if the cameras after the camera reconstruction step are inaccurate, as long as the epipolar geometry is accurate enough for the dense matching to work.

Acknowledgement The authors would like to thank Jana Kostková for her help in data acquisition and Martin Matoušek for his implementation of the rectification procedure. Tomáš Werner provided the routine for the bundle adjustment. The Valbonne images were provided by courtesy of Andrew Zisserman. This project is supported by a grant from the STINT Foundation in Sweden under Project Dur IG2003-2 062, by

the Czech Academy of Sciences under Project T101210406 and by IST Project IST-2001-39184 - BeNoGo.

References

1. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: BMVC. (2002) 384–393
2. Chum, O., Matas, J., Kittler, J.: Locally optimized RANSAC. In: DAGM. (2003) 236–243
3. Martinec, D., Pajdla, T.: Consistent multi-view reconstruction from epipolar geometries with outliers. In: SCIA. (2003) 493–500
4. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2000)
5. Fitzgibbon, A.: Simultaneous linear estimation of multiple view geometry and lens distortion. In: Proc. CVPR. Volume 1. (2001) 125–132
6. Šára, R.: Finding the largest unambiguous component of stereo matching. In: ECCV. (2002) 900–914
7. Kostková, J., Čech, J., Šára, R.: Dense stereomatching algorithm performance for view prediction and structure reconstruction. In: SCIA. (2003) 101–107
8. Šára, R.: Accurate natural surface reconstruction from polynocular stereo. In: Proc NATO Adv Res Workshop Confluence of Computer Vision and Computer Graphics. Number 84 in NATO Science Series, Kluwer (2000) 69–86
9. Šára, R., Bajcsy, R.: Fish-scales: Representing fuzzy manifolds. In: ICCV. (1998) 811–817
10. Faugeras, O., Keriven, R.: Complete dense stereovision using level set method. In: ECCV. (1998) 379–393
11. Kutulakos, K.N., Seitz, S.M.: A theory of shape by shape carving. IJCV **38** (2000) 199–218
12. Kolmogorov, V., Zabih, R.: Multi-camera scene reconstruction via graph cuts. In: ECCV. (2002) 82–96
13. Han, S., Medioni, G.: Reconstructing free-form surfaces from sparse data. In: ICPR. (1996) 100–104
14. Amenta, N., Bern, M., Kamvysselis, M.: A new Voronoi-based surface reconstruction algorithm. In: SIGGRAPH. (1998) 415–421
15. Kalaiah, A., Varshney, A.: Modeling and rendering of points with local geometry. IEEE Trans on Visualization and Computer Graphics **9** (2003) 30–42
16. Ferrari, V., Tuytelaars, T., van Gool, L.: Wide-baseline multiple-view correspondences. In: CVPR. (2003) I: 718–725
17. Zýka, V., Šára, R.: Polynocular image set consistency for local model verification. In: OeAGM Workshop. (2000) 81–88
18. Schaffalitzky, F., Zisserman, A.: Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In: ECCV. (2002) 414–431