GETTING TO GRIPS WITH 3D MODELING

Luc Van Gool[1,2,] Thomas Koninckx[1], and Tobias Jaeggli[1,2]

[1]ESAT / VISICS, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[2]D-ITET/BIWI, Swiss Federal Institute of Technology (ETH), Gloriastrasse 35, CH 8092, Zürich, Switzerland
phone: +32 16 321705, fax: +32 16 321723, email: {tkoninckx, vangool, tjaeggli}@esat.kuleuven.ac.be
web: www.esat.kuleuven.ac.be/psi/visics

ABSTRACT

3D acquisition technology has made big strides forward over the past years. Systems have become easier to use, cheaper, and faster. These developments are discussed in the plenary presentation that goes with this paper, both for the capture of shape and of surface textures. As the breadth of topics would only allow for a very superficial description of any particular method, the paper focuses on one such recent development as a good case in point: a structured light approach that supports interactive modeling. Textured 3D models are produced on-line, while manipulating the object in front of the system. This allows the user to check the quality of the result during the scanning and to perform effective view planning. Moreover, the projected patterns are automatically adapted to the scene. The system only requires a regular camera, LCD projector, and PC.

1. INTRODUCTION

The construction of 3D models is steadily becoming easier, cheaper, and faster. There are a number of technological trends behind these developments, which further accelerate their pace.

For one thing, several recent techniques are based on off-the-shelf hardware, not to say consumer products. Examples are structure-from-motion techniques that take uncalibrated video sequences from a handycam as input and turn them into 3D models of the circumnavigated objects. Similarly, several structured light techniques are based on readily available hardware components. In this paper one such example is discussed in more detail. That system is composed of a regular camera, an LCD projector, and a PC. This use of commodity products lets such systems benefit from the fast increase in quality and decrease in size and cost of cameras and projectors. As a result, while getting cheaper these systems also get better and more compact.

A second tendency is that 3D acquisition systems become more flexible. Recent systems are often portable and can be easily brought to the scene or object to be modeled, rather than v.v. Also, the range of object sizes they can handle is typically growing wider. Whereas different types of laser scanners serve different working volumes, several systems can now deal with that complete range.

Increasingly, 3D acquisition is also becoming an interactive, on-line process. Some video-based, passive systems yield the 3D structure and camera motion in real-time, while the video is being taken, even if the number of features dealt with is still limited. The active, structured light system discussed here has similar properties, but helped by the special illumination it already yields dense data.

In a similar vein, the number of images that need to be taken is decreasing. Rather than taking complete video streams as structure-from-motion input, a relatively small number of digital stills could be easier to take and offer better quality, due to their (much) higher resolution. In the case of structured light, the number of subsequent projections is steadily driven down, to minimize the overall capture time. One-shot systems get their 3D data from a single projection and image.

Future systems will need to better adapt to the object to be scanned. Now, the object or its environment is often adapted to the device, by carefully controlling the ambient light or by even powdering the object. Every technique has its limitations. Stereo-like systems require surface texture and will fail on homogeneously colored surfaces, laser scanners tend to fail on hair and similar fine structures and surface patterns may cause interference with the scanning, almost all systems have difficulties with specularities, etc. The future will see systems emerge that combine/switch between multiple reconstruction strategies, and apply each of these with automatically optimized parameters.

Another important evolution will be towards multiple units working simultaneously, in order to capture complete models in a very short time span. More importantly, this would then also allow such systems to capture dynamic scenes. Indeed, 3D acquisition technology has to a large extent been confined to the realm of static scenes. Capturing scenes in 3D with all the relative motions between objects is a challenge that so far requires the use of motion capture systems, which typically reduce the output to the trajectories of a limited number of special markers. By placing multiple cameras or structured light systems in the scene, dense and dynamic data can be acquired. Multi-camera rooms that capture visual hulls are a precursor to the more full-fletched implementations that will start to appear.

Acknowledgement: The authors gratefully acknowledge support by the K.U.Leuven Research Council GOA-project `MARVEL' and the Swiss SNF NCCR project `IM2'. Th. Koninckx is supported through a grant of the Flemish Institute for Scientific and Technological Innovation IWT.

The remainder of this paper describes one of the structured light developments we are currently working on, and which tries to take on these challenges. Whereas references to earlier work in general could not be given due to the immense volume of contributions, that particular effort will also be put into the context of the existing literature.

2. ON-LINE STRUCTURED LIGHT METHOD

From now on we focus on one particular development in our current work, namely a structured light approach intended to make active scanning cheaper, easier, and faster. Its main feature is that models are built on-line.

Cost is reduced by using off-the-shelf, consumer-grade components only. A PC, camera and an LCD projector are the only hardware required. Ease of use derives from the intuitive interaction with the system, where the model is built by letting the user manipulate the object in front of the scanner – hence the paper's title - and by providing immediate visual feedback. The latter obviously implies that the system should also be fast. This is achieved by combining two scanning modes. During manipulation the object is moving, and scanning is based on a single frame structured light method, which puts out 3D patches in real-time. This realtime data stream is presented on a monitor. As soon as the user is happy with the current pose, s/he keeps the object still. This triggers the system to switch to the second scanning mode. This produces higher quality 3D patches (range maps) based on a time sequence of projected patterns. Only these higher quality patches are integrated into the actual 3D model. The real-time patches streaming in during the single frame mode are shown aligned to the partially built up model. Of these, only the last patch is visualized, and this in a distinct color to indicate the current 'region of interest' on the object. This helps in keeping the overview of the parts already modeled and in deciding where the next multi-frame patch should be acquired in order to further complete the model. The visual feedback provided by the system is also crucial in checking the quality of the model, and to fill in holes immediately during the same scanning session, while this is still easy.

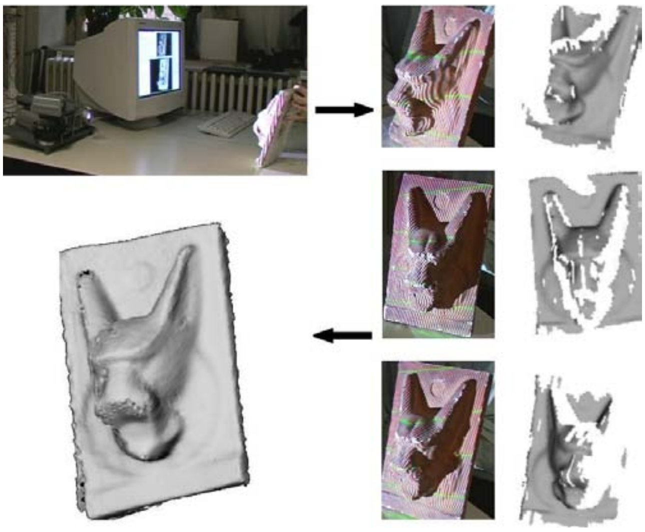

Figure 1: An overview of the acquisition and modelling pipeline. Top left: the setup while scanning a tablet with a dog head. Right: partial reconstructions coming from the scanner. Bottom left: an integrated model (shaded, no texture) based on approximately 600K triangles.

3. PREVIOUS WORK

There are two aspects of the proposed system where contenders have been rare. On the one hand, the system provides 3D data online, and also registers and visualizes these on the fly. On the other hand the system adapts its strategy to the scene content. We discuss these aspects in turn.

Although there is a substantial body of work on offline structured light techniques [1], relatively few provide their 3D output fast and online.

This said, the system presented here is not the first, nor the only one within its category. The system probably coming closest has been developed by Rusinkiewicz *et al.* [10]. Range data are obtained fast using a predefined, time-coded series of projection patterns. The acquisition speed allows for a slow motion of the scanned object. Surface texture is not captured. Popescu *et al.* have presented a hand-held scene modeling device, based on the projection of a fixed and rather sparse set of laser dots. This system also captures texture, but its resolution is still rather low, limiting its use to smooth surfaces. Tubic *et al.* [4] describe a similar hand held approach, providing higher resolution but no texture.

The resulting integrated surface is available at the moment of scanning, but the acquisition speed is limited because it only digitizes a small part of the surface per frame. As in our work and in that of Rusinkiewicz *et al.*, Blais *et al.* [5] also propose a system to which a hand-held object is presented. This powerful system uses fast autosynchronised laser-triangulation. Medium speed but very accurate acquisition puts the focus more on metrology and less on modeling. Another aspect of the work by Rusinkiewicz *et al.* and that presented in this paper, is that the acquired 3D patches are integrated on the fly.

The amount of work on structured light approaches that adapt their pattern(s) to the scene content has so far been more limited still. A color coded structured light approach (a continuation based on the initial work of [6]) in which the number of projection patterns and as such the noise margins are adapted to the actual noise level was proposed by Caspi et al. [7]. Related to this Horn et al. [8] proposed a technique to design an optimal sequential structured light pattern of length K. The work of Zhang *et al.* [9] is also based on color coding. They use a very nice dynamic programming based labeling to solve the correspondence problem in a single frame mode. Results however are only available offline. In case high-resolution scans are needed a scanning mode based on multiple images is used.

To the best of our knowledge, the proposed system is the first to combine self-adaptive and online structured light features.

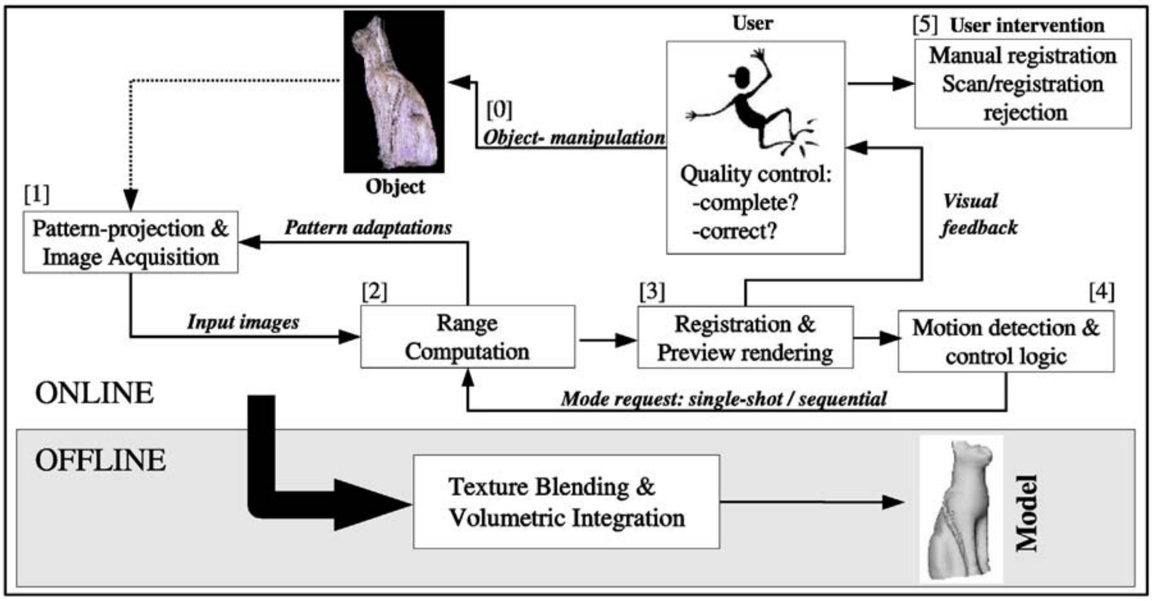
Object
[0]
Object- manipulation
User
[5] User intervention
Manual registration
Scan/registration rejection
Quality control:
-complete?
-correct?
Visual feedback
[1]
Pattern-projection & Image Acquisition
Pattern adaptations
[2]
Range Computation
Input images
[3]
Registration & Preview rendering
[4]
Motion detection & control logic
Mode request: single-shot / sequential
ONLINE
OFFLINE
Texture Blending & Volumetric Integration
Model

Figure 2: The 3D-Model Acquisition Pipeline, comprised of an online and an offline block. Online a crude model is built up and displayed on the fly in order to inform the user about the current status of the scanning process. Offline a high quality model is constructed.

4. MODEL ACQUISITION PIPELINE

Fig. 2 shows an overview of the proposed 3D modeling pipeline. Acquisition starts with presenting an object to the scanner (0). After grabbing an image with structured illumination (1) depth information is computed (2). Feedback from the range computation module to the pattern generation allows to optimize the pattern for the current object. The acquired surface patches are fed into a registration module (3), which presents the user with a preview of the partially reconstructed model.

The system automatically switches between two modes of operation based on motion detection (4):

- *tracking mode:* If the object is moving, range computation is based on real-time single frame acquisition. This allows to track the moving object and to update the estimate of its position. The last acquired scan is rendered in a different color to inform the user about the current view-angle with respect to the model. Subsequent scans are 'tracked' over the model built up so far. As these scans are possibly of lower quality, they are *not* used in the final model.
- *acquisition mode:* A high quality scan is automatically taken if the object stops moving, i.e. when kept still by the user. At that point a more accurate scan is made, based on a time series of projections. This scan is aligned and added to the model. The final model consists only of these high quality scans.

The scans of the first type are mainly intended for view planning and to avoid that the registration module would have to deal with large transformations between consecutive shots. As said, the second type of scans, are the ones used for the actual modeling. When enough such data have been acquired, the 3D textured model can then be refined in an extra, offline post-processing step, without user intervention.

In some situations, the user may wish to intervene manually during the online scanning (5), e.g. to remove individual scans. In addition, some objects are difficult to scan without interruption. An example is a thin planar object of which both front and back should be scanned. We won't be able to make a series of consecutive registrations which allow changing from the front to the back. Therefore the system offers the possibility to start a new sequence and register multiple partial 3D models later. The user can indicate such intentions by pushing a button on the PC's keyboard. Currently, the hands are either kept outside the field of view or are hidden by wearing black gloves. The user can artificially reduce the field of view by specifying a mask which invalidates part of the input image to remove disturbing structures in the environment.

4.1 Range acquisition

Range acquisition by the system is triangulation based. In the tracking mode, a single pattern is projected and range maps are generated at ca. 10 Hz. The pattern consists of black, parallel stripes on a white background. Colored 'code lines' transversal to the stripes yield a code for their unique identification. The structure of this pattern as well as the algorithm for the fast shape extraction are explained in [12]. There it is also described how the colors of the code lines, their number, as well as the width of the stripes are adapted online to the scene content. Here we need to introduce this pattern because of the role of the 'code points' – these are the points where the colored code lines intersect the stripes – in the decision to switch from tracking mode to acquisition mode. In the acquisition mode a traditional series of Gray encoded binary patterns [1] is projected.

The decision to switch between the modes is based on the following, simple motion detection algorithm. After the initial detection (see [12]) of the code points $x_{code,ij}$ with $i:1\rightarrow n$ the identifier of the generating code line $l_{code,i}$, and $j:1\rightarrow m$ the identifier of the corresponding stripe boundaries S_j we run the following predictor corrector style algorithm to keep track of these points over time:

- a predictor line $l_{predictor,i}$ is fitted to each set of code points $x'_{i,j}$ in the previous frame that belong to the same code line (i.e. regression over all intersections of one code line with the different stripes). Grouping of code points into such sets is a by-product of the stripe identification algorithm in [12]. Fitting a straight line is sufficient for our purposes.
- the intersections $l_{predictor,i} \bigcap S_j, \forall_j$ yield predicted code points $x''_{i,j}$ in the current frame. Their positions are refined to positions $x_{i,j}$ by convolving the image data around the position of S_j with the [5 × 7] environment of the projection pattern around the code point. This matched filter approach yields the new $x_{i,j}$ where the maximum response is found. The search is confined to a local neighborhood around the predicted positions.
- for every $x_{i,j}$ we search the closest match $x'_{i,j'}$ obeying the epipolar geometry between the current and the previous frame. The motion of a stripe S_j is limited given the frame-rate (between 10 and 20 fps dependent on scene complexity). This again allows the system to restrict the size of the search region.
- the matches between $x_{i,j}$ and $x'_{i,j'}$ are filtered by imposing that a consistent labeling between the current set of stripes is transferred in a consistent way to the previous labeling. Code points on the same stripe should remain on the same stripe in the previous frame.

The median distance between all $x_{i,j}$ and $x'_{i,j'}$ provides us with an inter-frame translation vector. This translation vector is fed into a Kalman filter, with 2D image position and motion as a state vector. The filter allows to distinguish between jitter and real motion, as a difference between random and systematic deviations from immobility.

At this point the system can decide when a safe transition from *tracking mode* to *acquisition mode* can be made. The constraint is that the median distance should be smaller than the stripe width. In this case reconstruction is not affected, otherwise strong artefacts will result. After returning from *acquisition mode* to *tracking mode* it is checked if the object was really immobile during the sequential acquisition. This validates the current scan.

4.2 Registration and rendering

The task of the *registration & rendering* module is to compute for each new scan the correct 6 DOF transformation with respect to the model-centered coordinate frame and render the currently accumulated scans as a preview of the final model.

In the following description, M_t refers to a range scan taken at time *t* and T_t refers to the corresponding transformation to bring this scan into the model-centered coordinate frame. The model-centered coordinate frame is chosen equal to the first scan's coordinate frame $(T_0 = I)$. At each time step t, the relative transformation T_t between an incoming scan M_t and the reference scan M_0 is computed. As the scanner is operating at a high frame-rate in *tracking mode*, changes in the transformation T between successive frames can assumed to be small. Therefore T_{t-1} can be used as an initial estimate of the transformation T_t. For each incoming frame, pair-wise alignment is performed with one of the previously accumulated high-quality scans, as follows:

- We select from the list of accumulated high quality scans, one scan M_i as the base scan for the alignment. We choose the scan M_i which minimizes $d(T_i, T_t)$, with $d(.,.)$ the distance between the camera centers corresponding to the two transformations.
- M_t is aligned with M_i using a fast variant of the ICP algorithm [10, 14, 15]. As T_{t-1} is used as initial estimate of T_t, the initial estimate for the pair-wise alignment is $(T_i)^{-1} * T_{t-1}$.
- Given the relative transformation T' between M_t and M_i as a result of the pair-wise alignment, T_t can be computed as $T_t = T_i * T'$.

The set of accumulated high-quality scans can be seen as a graph, with the scans representing the nodes. The relative transformations originating from pair-wise alignment represent the edges. As each incoming scan has been aligned with exactly *one* base-scan, this graph takes the form of a tree, with M_0 as a root. The fact that at each time-step *t* a new scan M_i is chosen to align, reduces accumulation of registration errors in the following way. Suppose that each incoming scan would only be aligned to the last acquired high quality scan. In this case, the tree becomes degenerate. It only contains a single branch of length $N - 1$, with N the number of acquired high-quality scans. The transformation of the (only) leave of this graph has been computed by concatenating $N - 1$ relative transformations. As each of these transformations have a finite accuracy, the transformation of the leave-scan will suffer from accumulation of errors. By selecting the scan M_i to align to, we are able to build a reasonably balanced tree rather than the described degenerate one. As the branches of a balanced tree have a length approx. ~ log(N), accumulation of errors is reduced a priori.

4.3 Offline texture and scan integration

The proposed online acquisition pipeline delivers a number of registered range scans. Our online registration focuses on speed rather than highest possible accuracy. Therefore the registration can be improved offline by performing a multi-view registration. Experience showed that quality of both range data coming from the *acquisition mode* and on line registration is high enough (see also results section) to yield good 3D models using standard post-processing techniques [16, 17, 18].

5. RESULTS

`Accuracy` : In order to get a good model out of the pipeline, we need:

- ’large scale’ accuracy: systematic errors on the scan data, render integration into a single model impossible.
- ’differential’ accuracy: resolution should be high enough to digitize fine geometric details.
- registration accuracy: the drift introduced by online registration should be smaller than what can be corrected during fine registration by an offline multi-view registration algorithm.

These conditions are checked by comparing known geometry with the data resulting from scanning. The same is done for registration.

The *systematic error* on a single scan is measured by digitizing two planes under a square angle. Both are covered by a checkerboard pattern which facilitates accurate localization in 3D-space of their vertices. This localization and the known geometry provide us with ’ground truth’ data. Figure 3 shows the result. For the black checks no reconstruction is made. The left column shows the scanned geometry (A) and the deviation from the ground truth data (B and color in A, B). The error distribution (fig. 3 C) and table 1 confirm that the scans are unbiased and mainly reside within 0.5 mm from the true geometry.[1] ’Strong textures’ have the tendency to introduce a slight increase in error as illustrated by the darker colors on the edges of the checks. This also explains a slightly higher noise level in this experiment than in the next one.

The *resolution* is checked by measuring a staircase with an exponential profile. The data of a random intersection is plotted in figure3 (D). A difference in height of 200 μm is visible. The first stair of 100 μm still introduces a visual artefact in the rendering, but is largely submerged under the noise level.

Registration was validated by making a sequence of which we know how the scans should align. By scanning an object rotating on a turntable, camera centers are known to lie on a circle and should be coplanar. This is shown in figure 3 (E,F). No prior knowledge about the motion was used in the estimation. The fact that the estimated trajectory of camera centers closes on itself is a strong indication that accumulation of registration errors is small.

`Models` : Figure 4 shows some models generated by our acquisition pipeline. All models are shown with and without texture. A reference photograph allows to evaluate the overall model quality. The ’dog tablet’ model is based on approximately 40 range and texture maps in acquisition mode, the ’potato-head’ on about 70, and the cat took 25. Given an average acquisition time of less than 0.5 sec for each high quality shot, the total scanning time remains limited.

mean error:	0.034 mm	*stand.deviation:*	0.16 mm
median error:	0.067 mm	*# samples:*	141503
max resolution:	200 μm		
focal length f:	16 mm	*base line D:*	300 mm
distance Z:	1100 mm		

Table 1: Accuracy of a single range scan

6. SUMMARY AND CONCLUSIONS

Technology to capture 3D object models is currently making fast progress. Several trends have been highlighted, that make these systems increasingly cheap, flexible, and fast. As a case in point, we have proposed a fast online model acquisition framework based on structured light. 3D data are generated, registered and visualized on the fly. This online behavior provides the user with visual feedback about the ongoing scanning session. The scanner uses line patterns which are automatically adapted to the scene. A second higher-level adaptation is controlled by motion detection, and switches the scanner from single-shot to sequential acquisition mode. The combination of a real-time method for view planning and a time coded approach renders this hybrid acquisition both versatile and robust. Ongoing research focuses on discarding the hands of the user when manipulating the object.

REFERENCES

[1] J.Batlle, E.Mouaddib and J.Salvi, Recent Pogress in Coded Structured Light as a Technique to Solve the Correspondence Prob: Survey, Pat. Recog. vol 31, nr.7,pp.963-982, 1998.

[3] V.Popescu, E.Sacks and G.Bahmutov, The modelcamera: a hand-held device for interactive modeling, 4’th Int. Conf.on 3-D Digital Imaging and Modeling, pp285-292, 2003.

[4] D.Tubic, P.Hebert, D.Laurendeau, 3D Surface Modeling from Curves, CVPR 2003,pp. I-842-849.

[5] F.Blais, M.Picard, G.Godin, Recursive Model Optimization Using ICP and Free Moving 3D Data Acquisition, 3DIM’03, pp251-259, 2003.

[6] K.Boyer, A.Kak, Color-encoded structured light for rapid active ranging, PAMI, Vol.9, pp.14-28, 1987.

[1] Given the current distance and volume.

Error $\sigma_Z \approx \frac{Z^2}{fD}\sigma_{\det ection}$

[7] D.Caspi, N.Kyriati, J.Shamir, Range Imaging With Adaptive Color Structured Light, PAMI, vol 20, nr.5,pp. 470-480, 1998.

[8] E. Horn, N. Kiryati, Towards Optimal Structured Light Patterns, Image and Vision Computing, Vol. 17, pp. 87-97, 1999.

[9] J.Zhang, B.Curless, S.M.Seitz, Rapid Shape Acquisition Using Color Structured Light andMulti-pass Dynamic Programming, 3DDPVT, Padova, Italy, June 2002.

[10] S. Rusinkiewicz, O. Hall-Holt and M. Levoy, Real-Time 3D Model Acquisition, SIGGRAPH, 2002, pp.438-446.

[12] T. Koninckx, A. Griesser, L. Van Gool, Real-time Range Scanning of Deformable Surfaces by Adaptively Coded Structured Light, Fourth International Conference on 3-D Digital Imaging and Modeling (3DIM03), S. Kawada, ed., p. 293-302, 2003, IEEE Computer Society.

[13] A.McIvor, R.Valkenburg, Calibrating a structured light system, Image & Vision Computing New Zealand, pp. 167-172, Industrial Research Limited, 1995.

[14] P.J Besl and N.D. McKay, A method for registration of 3-d shapes. PAMI, 14(2):239-256, February 1992.

[15] Y. Chen and G. Medioni, Object modelling by registration of multiple range images. *Image and Vision Computing* 10(3) :145-155, April 1992.

[16] K. Pulli, Multiview Registration for Large Data Sets, 3DIM'99, Ottawa, pp.160-168, 1999.

[17] B. Curless, M. Levoy, A Volumetric Method for Building Complex Models from Range Images. SIGGRAPH, 1996, p.303-312.

[18] J. Davis, S.R.Marschner, M.Garr, M.Levoy, Filling holes in complex surfaces using volumetric diffusion, First Int. Symp. on 3D Data Proc., Vis., and Transmission, 2002.

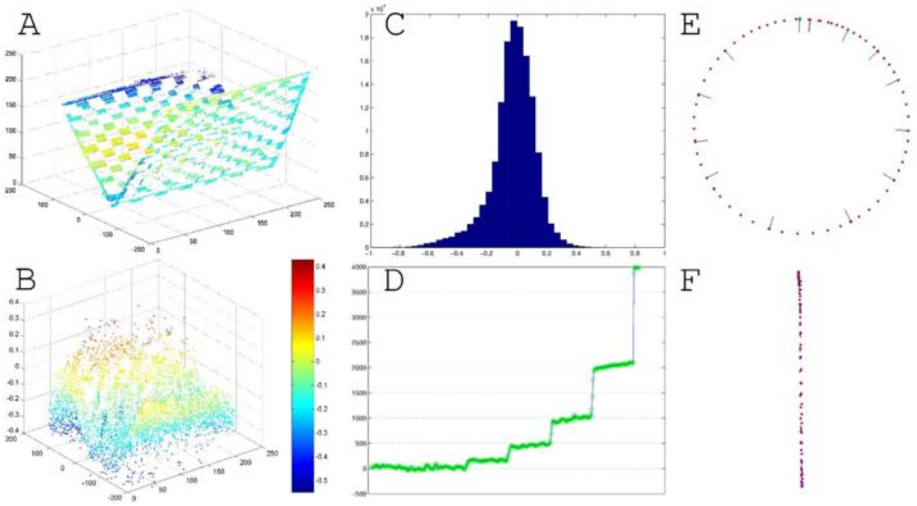

Figure 3: Accuracy. [A]: geometry plot with deviation from ground truth data indicated in color (mm). [B]: absolute error for the geometry shown in A. Color is consistent between A and B. [C]: error distribution. [D]: resolution in μm [E,F]: error on registration. The camera centers should make a perfect circle (top) and should be coplanar (bottom). Tracked cameras (real-time data) are shown as dots, high quality shots as lines.

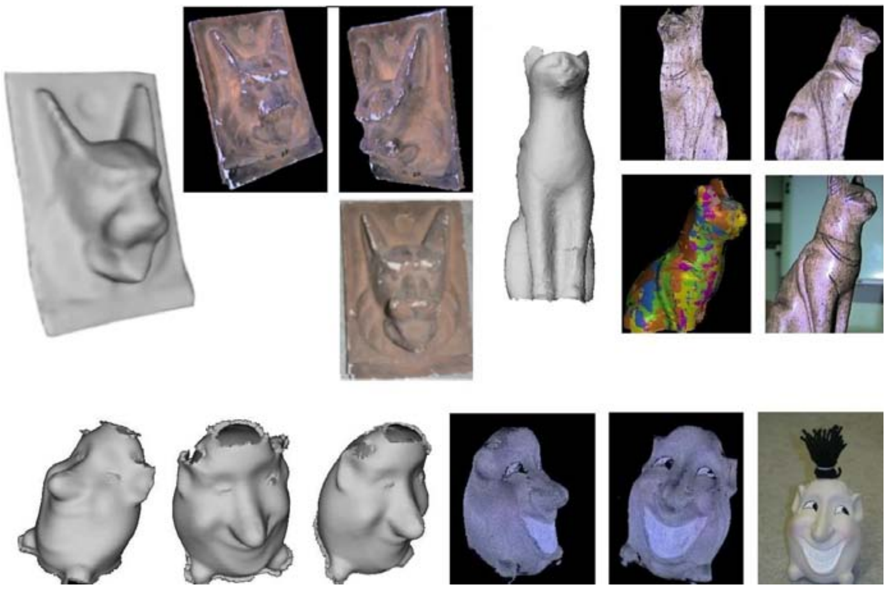

Figure 4: A collection of objects digitized with our system. Top left: a stone tablet with a dog head. Top right: a model of a glossy statue of a cat. Bottom: a 'potato-head' figurine. Remark: the hole on top is due to the 'hair' which can't be scanned, the holes near the ears are part of the geometry. With every group, the image in the bottom right is a reference photograph. The colored version of the cat shows the registration of the individual patches.