Live 3D Shape Reconstruction, Recognition and Registration

*Carlos Hernández[2†] *Frank Perbet[1] *Minh-Tri Pham[1] *George Vogiatzis[3†] *Oliver J. Woodford[1]
Atsuto Maki[1] Björn Stenger[1] Roberto Cipolla[4]
[1]Cambridge Research Laboratory, Toshiba Research Europe Ltd, Cambridge, UK
[2]Google, Seattle, US [3]Computer Science, Aston University, Birmingham, UK
[4]Department of Engineering, University of Cambridge, Cambridge, UK

We present a video-based system which interactively captures the geometry of a 3D object in the form of a point cloud, then recognizes and registers known objects in this point cloud in a matter of seconds (fig. 1). In order to achieve interactive speed, we exploit both efficient inference algorithms and parallel computation, often on a GPU. The system can be broken down into two distinct phases: geometry capture, and object inference. We now discuss these in further detail.

Geometry capture

The reconstruction phase consists of two key steps: pose estimation of each video frame, and dense geometry estimation using the input frames and their computed poses. We have two interchangeable methods for real-time camera pose estimation. Both assume known internal camera parameters. The first is PTAM [1]—an off-the-shelf visual SLAM algorithm. The second computes pose metrically (required here for object inference) using a known planar pattern [5] (fig. 2(a)), as well as being fast and robust.

Given pose, computing points by finding frame-to-frame correspondences becomes a 1D search (fig. 2(b)) (assuming a static scene). Even so, accurately matching hundreds of thousands of correspondences over multiple frames can be computationally expensive if approached naively. We use a probabilistic framework [3] which maintains a very compact state per correspondence over time. Further speed-up is achieved by matching, using NCC on 5×5 windows, on a GPU. Note that each point is computed independently—there is no regularization.

Object inference

Recognition and registration is done jointly, in a phase consisting of four key steps [2]. The first step converts the point cloud to a 128^3 voxel volume (fig. 3(a)) using a Gaussian on the distance of each voxel centre to the nearest point.

The second step finds features over scale and translation using the DoG detector, then computes orientation using PCA on the surrounding volume, to generate a local, 7D feature pose (fig. 3(b)). A simple descriptor is computed by sampling the volume (at the correct scale) at 31 regularly distributed locations around the feature point. The entire feature extraction pipeline is implemented on a GPU.

*Joint first authors. †Work done while at Toshiba Research Europe Ltd.

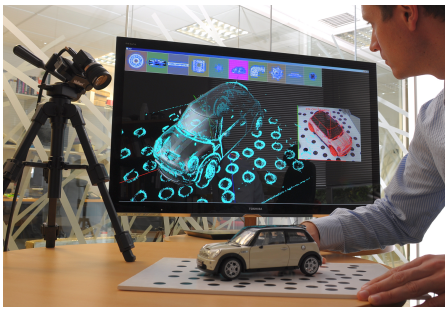

Figure 1. **Our system** in use. The user rotates the object on the pattern, and it is reconstructed on screen (cyan dots). After object inference the recognized objects are overlaid on the point cloud and on the video output.

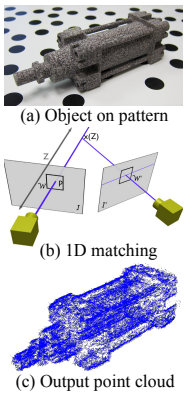

Figure 2. **Geometry capture** key steps.

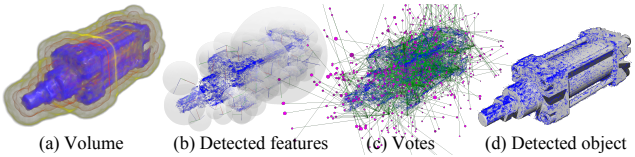

Figure 3. **Object inference** key steps.

In the third step, 8D votes over object class and pose are computed (fig. 3(c)) by matching features with those extracted from training data of known class and pose.

In the final step, modes of density in the vote space indicating the presence of an object are found using the minimum-entropy Hough transform [4], which additionally explains away incorrect votes. The density is defined using a scale-invariant distance between votes [2].

Additional physical constraints, such as object size, height off the ground and collision detection, can be used to filter the list of detected objects further.

References

[1] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, 2007. 1

[2] M.-T. Pham, O. J. Woodford, F. Perbet, A. Maki, B. Stenger, and R. Cipolla. A new distance for scale-invariant 3D shape recognition and registration. 2011. 1

[3] G. Vogiatzis and C. Hernández. Video-based, real-time multi view stereo. *Image and Vision Computing*, 29(7):434–441, 2011. 1

[4] O. J. Woodford, M.-T. Pham, A. Maki, F. Perbet, and B. Stenger. Demisting the Hough transform for 3D shape recognition and registration. In *BMVC*, 2011. 1

[5] Z. Zhang. A flexible new technique for camera calibration. *TPAMI*, 22(11):1330–1334, 2000. 1