Live Metric 3D Reconstruction on Mobile Phones

Petri Tanskanen, Kalin Kolev, Lorenz Meier, Federico Camposeco, Olivier Saurer, Marc Pollefeys

ETH Zurich, Switzerland

ICCV 2013
Main Contents

1. Target & Related Work
2. Main Features of This System
3. System Overview & Workflow
4. Detail of This System
5. Experiments
6. Conclusion
1. Target & Related Work

• The first **dense** stereo-based system for live **interactive** 3D reconstruction on mobile phones. It generates dense 3D models with **absolute scale** on-site while simultaneously supplying the user with real time interactive feedback.
Related work

- Wendel et al.\cite{1} rely on a distributed framework with a variant of PTAM on a micro air vehicle. All demanding computations are performed on a separate server machine that provides visual feedback to a tablet computer.

![Diagram of distributed reconstruction system.]  

Figure 2. Overview of the distributed live dense reconstruction system, which exploits the individual capabilities and requirements of the system entities. The tracker is run on the quad-rotor, sparse and dense mapping on the server, and visualization on the tablet.

\[1\] A. Wendel, M. Maurer, G. Graber, T. Pock, and H. Bischof. Dense reconstruction on-the-fly. CVPR 2012.  
( Graz University of Technology, Austria )
Pan et al.\cite{2} demonstrated an interactive system for 3D reconstruction on a mobile phone.

\cite{2} Q. Pan, C. Arth, E. Rosten, G. Reitmayr, and T. Drummond. Rapid scene reconstruction on mobile phones from panoramic images. ISMAR, 2011. (Cambridge University & Graz University of Technology)
Related work

- Pisacariu et al.\cite{3} presented a \textit{shape-from-silhouette framework} running in real time on mobile phone.

\cite{3} V. A. Prisacariu, O. Kaehler, D. Murray, and I. Reid. Simultaneous. 3D tracking and reconstruction on a mobile phone. ISMAR, 2013. (University of Oxford)
2. Main Features of This System

(1) **Initialization**: fully automatic; markers or any other specific settings are not required.

(2) **Estimate the metric scale of the reconstructed 3D models**: feature-based tracking and mapping in real time; inertial sensing in position and orientation to estimate the metric scale of the reconstructed 3D models.

(3) **Interactive**: automatically select suitable keyframes when the phone is held still; use the intermediate motion to calculate scale; Visual and auditory feedback is provided to enable intuitive and fool-proof operation.

(4) **Dense stereo matching**: an efficient and accurate multi-resolution scheme for dense stereo matching and GPU acceleration; reduce the processing time to interactive speed.
The example of how this system works:

- Demo
3. System Overview & Workflow

Two main input streams of this system:
(1) camera frames: 640*480; 15-30 Hz
(2) inertial sensor information:
   angular velocity: 200Hz
   linear acceleration: 100Hz

The output is a 3D model in metric coordinates in form of a colored point cloud.

This system consists of three main blocks:
Inertial tracking; visual pose estimation; Dense 3D modeling.
3. System Overview & Workflow

$R_B$ is the rotation from the current world to body/camera frame.

$R_v$ is the rotation refinement with the visual tracker.

$x_f$, $x_v$, $x_i$ donate the fused, vision and inertial position estimates in the world coordinate.
4. Detail of This System

- 4.1 Initialization
- 4.2 Inertial Sensor
- 4.3 Visual Tracker
- 4.4 Sparse Mapping
- 4.5 Depth 3D Modeling
4.1 Initialization

- (1) Two View Initialization: The map is initialized from two keyframes.

Keyframe-1

Keyframe-2 (the inertial estimator detects a salient motion with a minimal baseline)

ORB features

ORB features extracted and matched

RANSAC + 5-point algorithm

Relative pose (R, t)

Matched points are triangulated
4.1 Initialization

- (2) A denser initial map:

```
Fast corners extracted and 8*8 patch as descriptor
```

```
Compare the ZSSD value along the segment of the epipolar line
```

```
Matched points are triangulated
```

```
Included to the map & bundle adjustment
```

```
Rotate the map
```
4.1 Initialization

- (2) A denser initial map:

- Fast corners extracted and 8*8 patch as descriptor
- Compare the ZSSD value along the segment of the epipolar line
- Matched points are triangulated
- Included to the map & bundle adjustment
- Rotate the map
4.1 Initialization

The camera and IMU are considered to be at the same location and with the same orientation.

\( R_B \) is the rotation from the current world to body/camera frame.

\[
R_B = [r_{xB}, r_{yB}, r_{zB}] \in SO(3)
\]

\[
r_{zB} = \frac{g_B}{\|g_B\|}, \quad r_{yB} = \frac{r_{zB} \times m_B}{\|r_{zB} \times m_B\|}, \quad r_{xB} = r_{yB} \times r_{zB}, \quad m_B \text{ is the visual measurements.}
\]
Introduction of the Inertial Sensors

- Inertial Measurement Unit (IMU): gyroscope and accelerometer
An accelerometer is a sensor for testing the acceleration along a given axis.

When a physical body accelerates at a certain direction, it becomes subject to a force equal to: \( F = ma \) in accordance with Newton's Second Law.

Output: Provide the three components of the acceleration on the three directions under the coordinate system defined by the device.
Gyroscope

A gyroscope is a device for measuring or maintaining orientation, based on the principles of angular momentum:

When no external torque acts on an object or a closed system of objects, no change of angular momentum can occur.

Output: provides the three components of the angular velocity under the coordinate system defined by the device.
System workflow

Initialization

Visual tracker

Inertial Sensors

When scale is fixed

Sparse Mapping

Depth 3D Modeling


3.2 Inertial Sensor

Pose Prediction with the Inertial Sensor:

When scale is fixed

Inertial Sensors

Visual tracker

Kalman Filter

Gyroscope

Verlet Integration

Accelerometer

Integration

$R_B$

$x_v$

$x_i$

$x_f$

$\mathbf{a}_B - g_B$

When scale is fixed
4.2 Inertial Sensor

The estimation of the rotation: \( \mathbf{R}_B \)

The filter prediction and update equations

\[
\hat{\mathbf{R}}_B = e^{\mathbf{ω}_B \Delta \mathbf{t}} \mathbf{R}_B^-,
\]

\[
\mathbf{r}^+_i B = \hat{\mathbf{r}}_i B + L_i k (z_i - \hat{\mathbf{r}}_i B) \text{ with } i \in (x, y, z),
\]

\[
\dot{\mathbf{R}}_B = \mathbf{ω} \mathbf{R}_B
\]

\[
\mathbf{r}_{zB} = \frac{\mathbf{g}_B}{\|\mathbf{g}_B\|}, \quad \mathbf{r}_{yB} = \frac{\mathbf{r}_{zB} \times \mathbf{m}_B}{\|\mathbf{r}_{zB} \times \mathbf{m}_B\|}, \quad \mathbf{r}_{xB} = \mathbf{r}_{yB} \times \mathbf{r}_{zB},
\]
4.2 Inertial Sensor

The estimation of the positions

\[
\vec{v}_I^{k+1} = \vec{v}_I^k + \tau \Delta t R_B \left( \vec{a}_B^k - g_B \right)
\]

\[
x_f = \kappa \left( \sigma_v^{-2} \lambda \vec{v}_v + \sigma_i^{-2} \vec{x}_i \right)
\]

\( f, v \) and \( i \) denote fused, vision and inertial position estimates, \( k \) is the normalizing factor.
3.2 Inertial Sensor
Metric scale estimation with the inertial sensors

The scale for visual-inertial fusion:

\[ \vec{x}_i = \lambda \vec{y}_i \]

\( \vec{x}_i \) : the displacement estimated by accelerometer

\( \vec{y}_i \) : the displacement estimated by vision
Metric scale estimation with the inertial sensors

- In order to deal with the noise and time-dependent bias from the accelerometer, an event-based outlier-rejection scheme is proposed.

\[ \| \vec{x}_i - \lambda \vec{y}_i \| > \text{threshold} \]

Find the optimal scale: given \((\vec{x}_i, \vec{y}_i)\)

\[ \arg\min_{\lambda} = \sum_{i \in I} \| \vec{x}_i - \lambda \vec{y}_i \|^2. \]
Metric scale estimation with the inertial sensors

As soon as the scale estimation converges, we can update the inertial position with visual measurements.
4.2 Inertial Sensor

Figure 4. Visual inertial pose estimate vs. ground truth.
System workflow

Initialization

Visual tracker

Inertial Sensors

When scale is fixed

$R_B$

$R_v$

$x_i$

$R_f$

$x_f$

Sparse Mapping

Depth 3D Modeling

When scale is fixed
4.3 Visual Tracker

- Refine the pose estimate from the inertial pose estimator and correct drift.
If the visual tracking is lost, image localization module from PTAM is used.
System workflow

 Initialization

 Visual tracker

 Inertial Sensors

 Sparse Mapping

 Depth 3D Modeling

 When scale is fixed

 $R_v$

 $x_v$

 $x_f$

 $R_B$

 $x_i$
4.4 Sparse mapping

Add new map points:

- New keyframes: moved the camera a certain amount; or the inertial position estimator detects that the phone is held still after salient motion.

- A list of candidates of the new map points:
  non maximum suppressed FAST corners + Shi-Tomasi score > a certain threshold.
Create a mask indicate the already covered regions: Overcome to map the already exist points.
4.4 Sparse Mapping

Priority:   Local Bundle Adjustment > Keyframes optimization for dense modeling > Global bundle adjustment

After a keyframe is added,
System workflow

Initialization

Visual tracker

Inertial Sensors

When scale is fixed

Depth 3D Modeling

Sparse Mapping

Rv

Rv

xf

xi

Rb
The core of the 3D modeling module is a stereo-based reconstruction pipeline.

- Image mask estimation
- Depth map computation
- Depth map filtering
4.5 Dense 3D Modeling

- Image Mask Estimation: sufficient material texture region and covered by the current point cloud regions

  (1) A texture-based mask

  (2) A coverage mask

  Depth map computations are restricted to pixels within the mask.
4.5 Dense 3D Modeling

texture-based mask

Shi-Tomasi Score > $\lambda_{min}$

Area covered by map points

coverage mask
4.5 Dense 3D Modeling

Depth Map Computation:

- Binocular stereo: an incoming image as a reference view and matching it with an appropriate recent image in the provided series of keyframes.
- A multi-resolution scheme.

Flowchart:

1. downsampling the input images
2. estimating depths
3. upgrading & refining the results
• An update scheme based on the current downsampled pixel position and three appropriate neighbors.
4.5 Dense 3D Modeling

Depth range

\[
\min \{ D^l_{i+1} | l = 0, \ldots, 3 \} \\
\max \{ D^l_{i+1} | l = 0, \ldots, 3 \}
\]
4.5 Dense 3D Modeling

The multi-resolution approach: 5 times faster than the single-resolution approach.

Figure 5. Single- vs. multi-resolution depth map estimation. From left to right: The reference image of a stereo pair, corresponding depth map estimated with a classical single-resolution winner-takes-all strategy and result obtained with the proposed multi-resolution scheme.
4.5 Dense 3D Modeling

GPU acceleration

• Parallelization potential of the algorithm with a GPU implementation.
• Reduce the overall runtime of the 3D modeling module to about 2-3 seconds per processed image.
Image Pair Selection

A crucial step in binocular stereo is the choice of an appropriate image pair.

An ideal candidate pair should share a large common field of view, a small but not too small baseline and similar orientations.
Image Pair Selection

Keyframe j

\[ C(j, k) = \cos \theta_{\text{pose}}^{jk} \cdot \cos \theta_{\text{view}}^{jk} \cdot \cos \theta_{\text{up}}^{jk} \]

Depth map

Stereo matching

j-1

j-2

j-3

j-4

j-5
4.5 Dense 3D Modeling

Depth map filtering: remove virtually all outliers and build a clean 3D model.

- Check the consistency over multiple views: \( |\hat{d}_i - d_i| \)

\( d_i \) is the value stored in the depth map.

\( \hat{d}_i \) is the depth value computed from other depth map.

\( |\hat{d}_i - d_i| < \delta \) in \( N_C \) views

The depth of X is considered consistent.
5. Experiments

• Platform: Samsung Galaxy SIII I9300GT with Samsung Exynos 4 quad core CPU and ARM Mali-400 MP4 GPU.

• Processed in real time.
Experimental Results

Non-movable objects for which no 3D geometry exists yet.

Figure 6. Photo and front and right view of the reconstructed 3D model of a 0.5 m tall African tribal mask.

Figure 7. Photo and front and left views of reconstructed 3D model of a 1.6 m tall Shakyamuni Buddha statue.

Captured from the collection of a museum
Experimental Results

Generality of the approach

Outdoor environments

Human faces

Figure 8. Photo and reconstructed 3D model of a building facade captured at street-level.

Figure 9. Front and left view of a reconstructed 3D model of a human face, including a corresponding photo of the test person.
6. Conclusion

- The first interactive on-device system for dense stereo-based 3D reconstruction on mobile phones.

- **Inertial sensors**: improve the resilience of the camera tracking process to rapid motions; automatically capture keyframes when the phone is static; derive the **metric measures** of the captured scene.

- An efficient and accurate method for **binocular stereo** based on a **multi-resolution** scheme.
• Thanks for your suggestion!