Unification des approches basées primitives et directes pour la localisation temps réel avec des capteurs RGB-D

Andrew Comport

I3S/CNRS-Université de Nice Sophia-Antipolis

Journées Nationales de la Recherche en Robotique 2015

22 octobre 2015

Context: Real-time dense localisation and mapping

RGB-D sensor

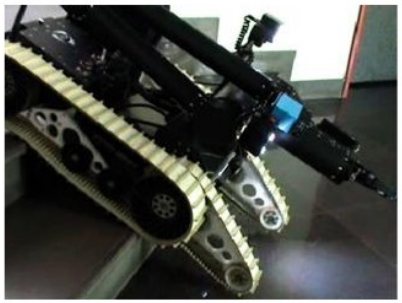
- Microsoft kinect, Asus Xtion, Intel Realsense, Orsens, standard stereo camera ...

Autonomous navigation

- Mapping/Exploration of uneven terrain

Hand-held scanning

- User feedback



ECA Robotics

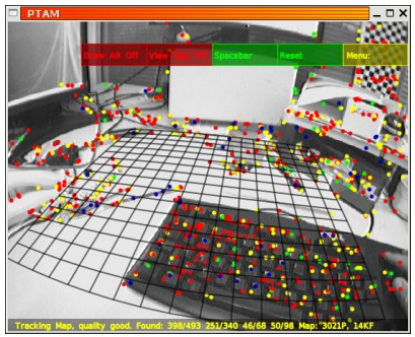
CNRS-I3S UNS

CNRS-AIST JRL

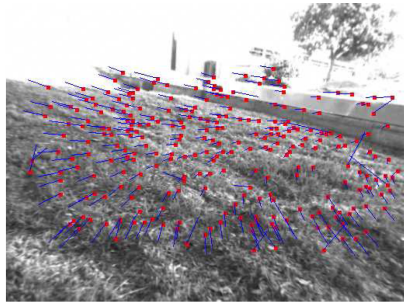
Advances in real-time dense RGB-D visual representations

Feature based approaches

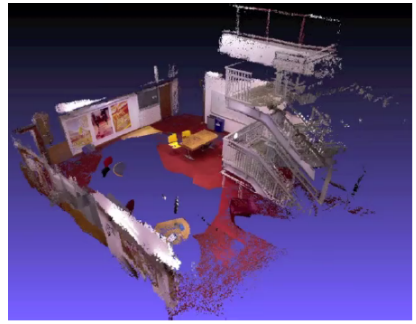
Multi-view stereo: GPS/INS-only pose estimation Mordohai et al. (2007)



PTAM Klein and Murray (2007)

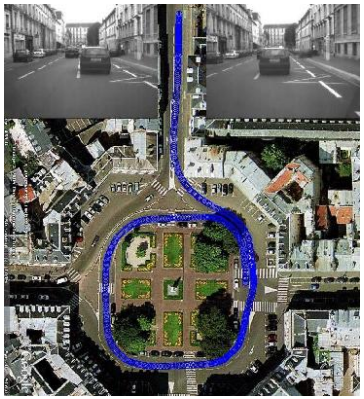


Outdoor mapping and Navigation using Stereo Vision Konolige et al. (2007)

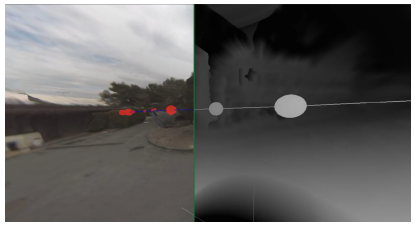


RGB-D Indoor Mapping Henry et al. (2010)

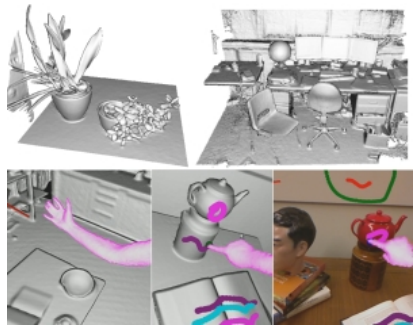
Direct approaches



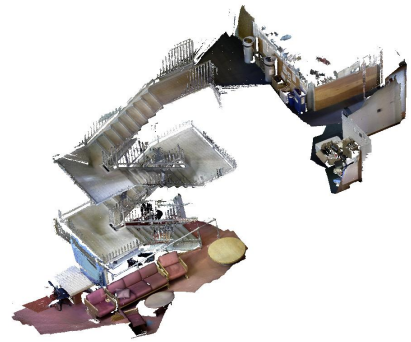
Dense stereo key-frame mapping and pose estimation, Comport et al. (2007)



Dense omni-directional key-frames Meilland et al. (2010)



Kinect fusion, volumetric, Newcombe et al. (2011)

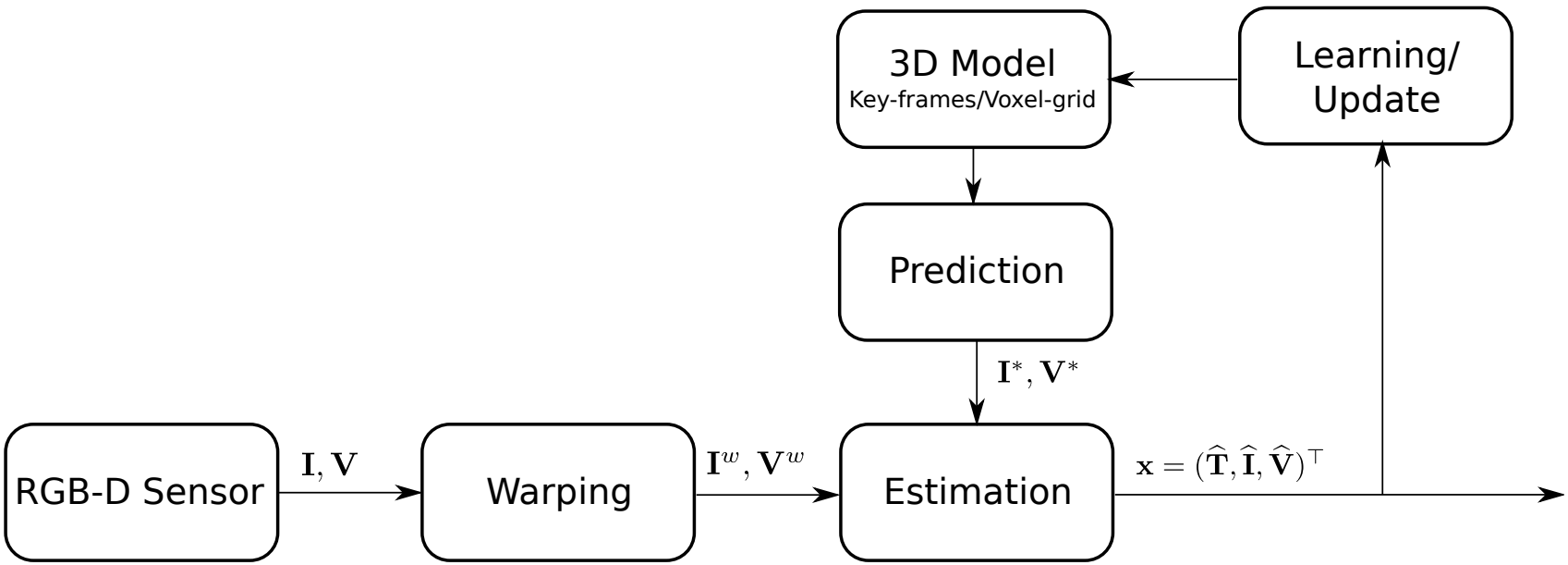


Kintinuous, volumetric + rolling buffer, Whelan et al. (2013)

Standard Tracking Pipeline

Standard tracking pipeline

- Sensor pose estimation
- Model update/integration
- Next view prediction



Pose Estimation

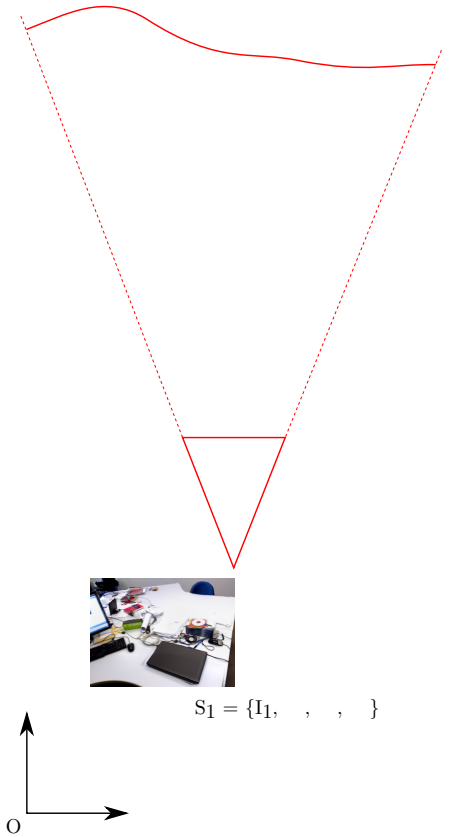
Key-frame $\mathbf{S}$ (local map)

- RGB values, $\mathbf{I} : n \times m \times 3$
- Depthmap, $\mathbf{D} : n \times m$
- Pose, $\mathbf{T} \in \mathbb{SE}(3)$
- Uncertainties, $\mathbf{C} : n \times m$

(Khoshelham et al. 2012)

Graph (global map)

- Multiple key-frames Meilland and Comport (2013a)
- Partial overlap



Pose Estimation

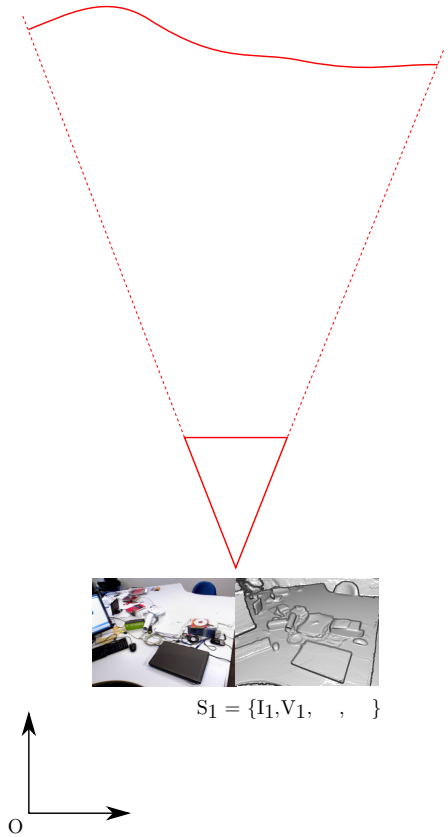
Key-frame $\mathbf{S}$ (**local map**)

- RGB values, $\mathbf{I} : n \times m \times 3$
- Depthmap, $\mathbf{D} : n \times m$
- Pose, $\mathbf{T} \in \mathbb{SE}(3)$
- Uncertainties, $\mathbf{C} : n \times m$

(Khoshelham et al. 2012)

Graph (**global map**)

- Multiple key-frames Meilland and Comport (2013a)
- Partial overlap



Pose Estimation

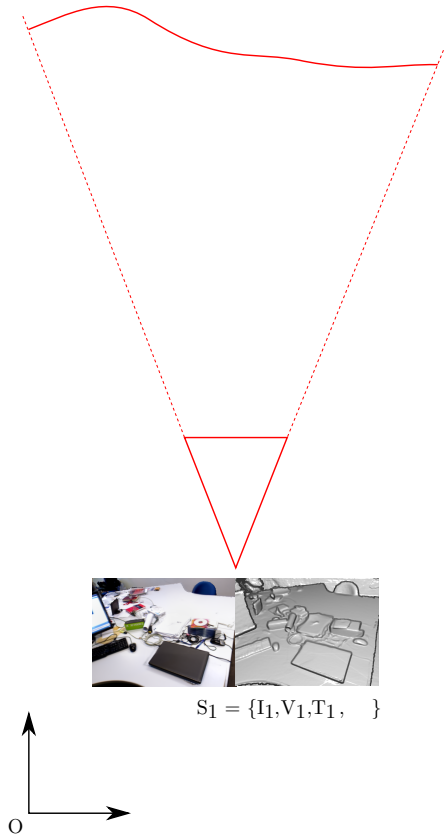
Key-frame $\mathbf{S}$ (local map)

- RGB values, $\mathbf{I} : n \times m \times 3$
- Depthmap, $\mathbf{D} : n \times m$
- Pose, $\mathbf{T} \in \mathbb{SE}(3)$
- Uncertainties, $\mathbf{C} : n \times m$

(Khoshelham et al. 2012)

Graph (global map)

- Multiple key-frames Meilland and Comport (2013a)
- Partial overlap



Pose Estimation

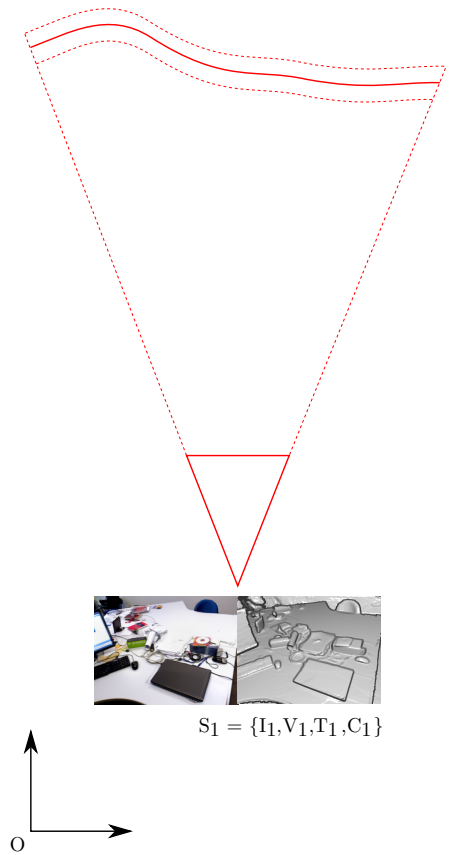
Key-frame $\mathbf{S}$ (local map)

- RGB values, $\mathbf{I} : n \times m \times 3$
- Depthmap, $\mathbf{D} : n \times m$
- Pose, $\mathbf{T} \in \mathbb{SE}(3)$
- Uncertainties, $\mathbf{C} : n \times m$

(Khoshelham et al. 2012)

Graph (global map)

- Multiple key-frames Meilland and Comport (2013a)
- Partial overlap



Pose Estimation

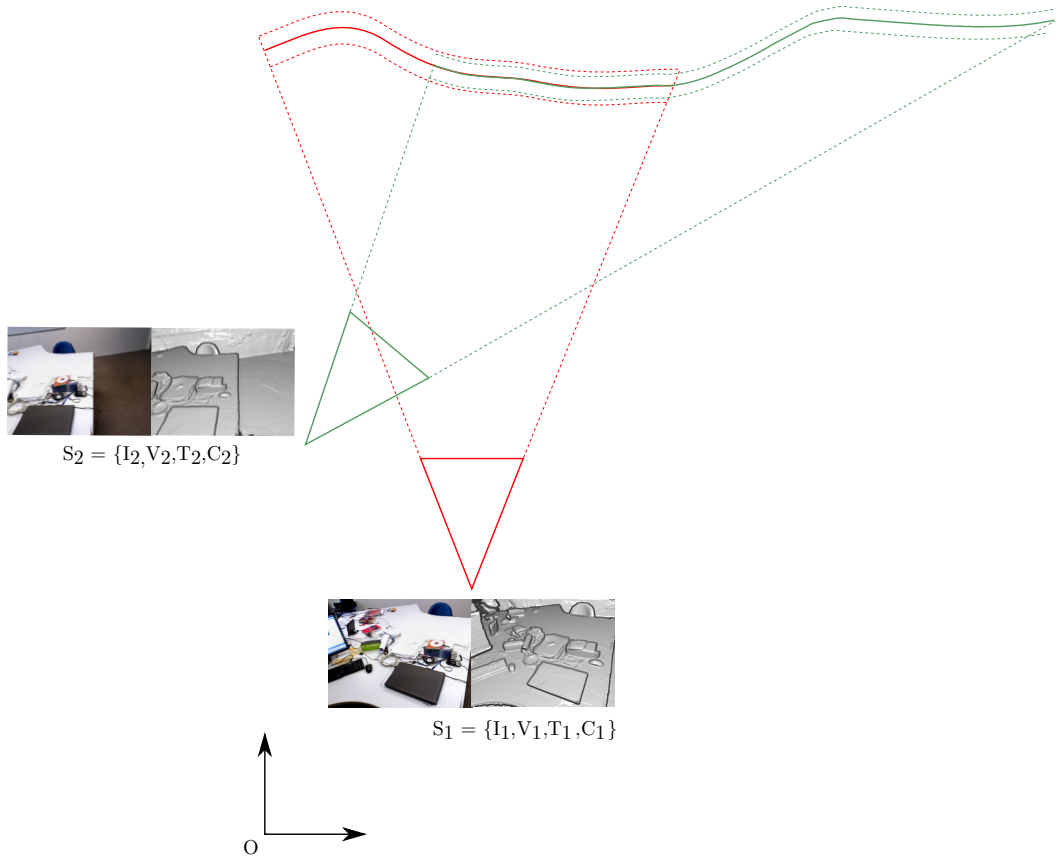
Key-frame $\mathbf{S}$ (**local map**)

- RGB values, $\mathbf{I} : n \times m \times 3$
- Depthmap, $\mathbf{D} : n \times m$
- Pose, $\mathbf{T} \in \mathbb{SE}(3)$
- Uncertainties, $\mathbf{C} : n \times m$

(Khoshelham et al. 2012)

Graph (**global map**)

- Multiple key-frames Meilland and Comport (2013a)
- Partial overlap



Pose Estimation

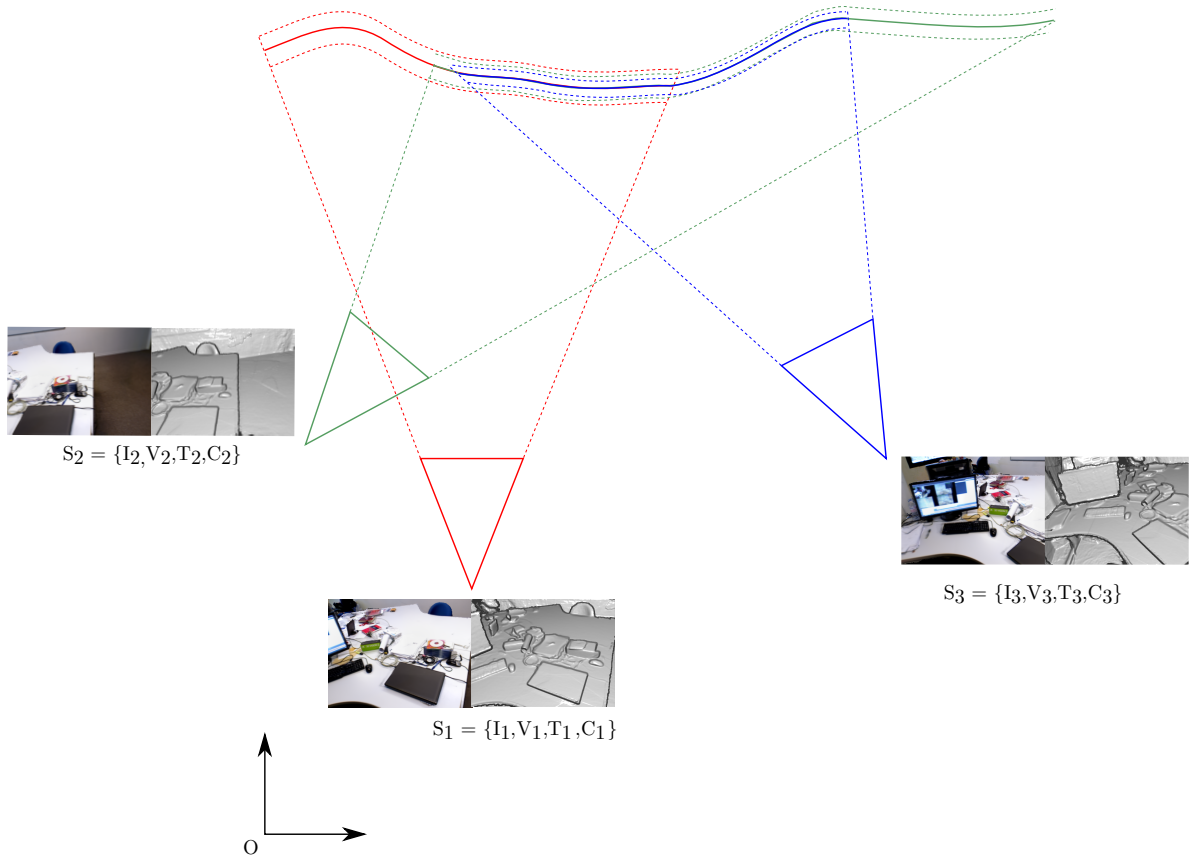
Key-frame $\mathbf{S}$ (**local map**)

- RGB values, $\mathbf{I} : n \times m \times 3$
- Depthmap, $\mathbf{D} : n \times m$
- Pose, $\mathbf{T} \in \mathbb{SE}(3)$
- Uncertainties, $\mathbf{C} : n \times m$

(Khoshelham et al. 2012)

Graph (**global map**)

- Multiple key-frames Meilland and Comport (2013a)
- Partial overlap



Iterative Closest Point (ICP) pose estimation

3D point cloud registration : Non-linear registration of two point clouds

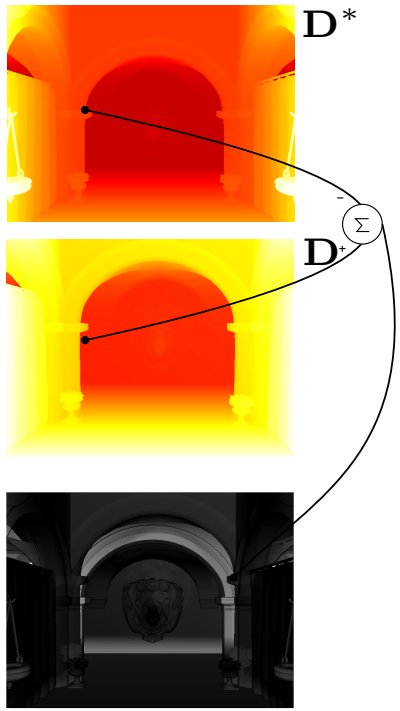
Point-to-Point ICP

$$\mathbf{e_D}(\mathbf{x}) = \mathbf{v} - \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})\mathbf{v}^*$$

$$\mathbf{x} = \arg\min_{\mathbf{x}} \mathbf{e_D}^T \mathbf{W_D} \mathbf{e_D}$$

Point-to-Point ICP

- ✓ Texture-less areas
- ✓ Convergence domain
- ✗ Noise sensitive (bilateral filtering)
- ✗ Symmetry $\rightarrow$ under-contrained estimation



$\mathbf{e_D}(\mathbf{x})$

- $\mathbf{x}$ is the 6dof pose unknown related to the homogeneous matrix $\mathbf{T} \in \mathbb{SE}(3)$
- $\mathbf{v}$ are the 3D verticies

Iterative Closest Point (ICP) pose estimation

3D point cloud registration : Non-linear registration of two point clouds

Point-to-Plane ICP

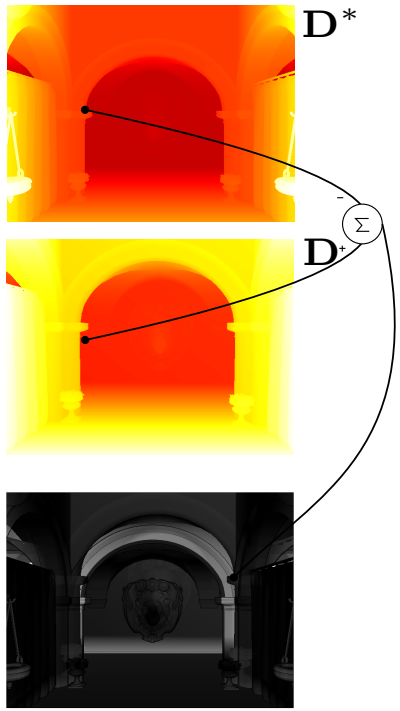
$$\mathbf{e_D}(\mathbf{x}) = \mathbf{N}(\mathbf{p})^T(\mathbf{v} - \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})\mathbf{v}^*)$$

$$\mathbf{x} = \arg\min_{\mathbf{x}} \mathbf{e_D}^T \mathbf{W_D} \mathbf{e_D}$$

Point-to-Plance ICP

- ✓ Texture-less areas
- ✓ Efficiency and Convergence domain
- ✗ Noise sensitive (bilateral filtering)
- ✗ Unconstrained scenes

- $\mathbf{N}$ contains the surface normals for each vertex



$\mathbf{e_D}(\mathbf{x})$

Feature-based image registration - pose estimation

Feature-based registration

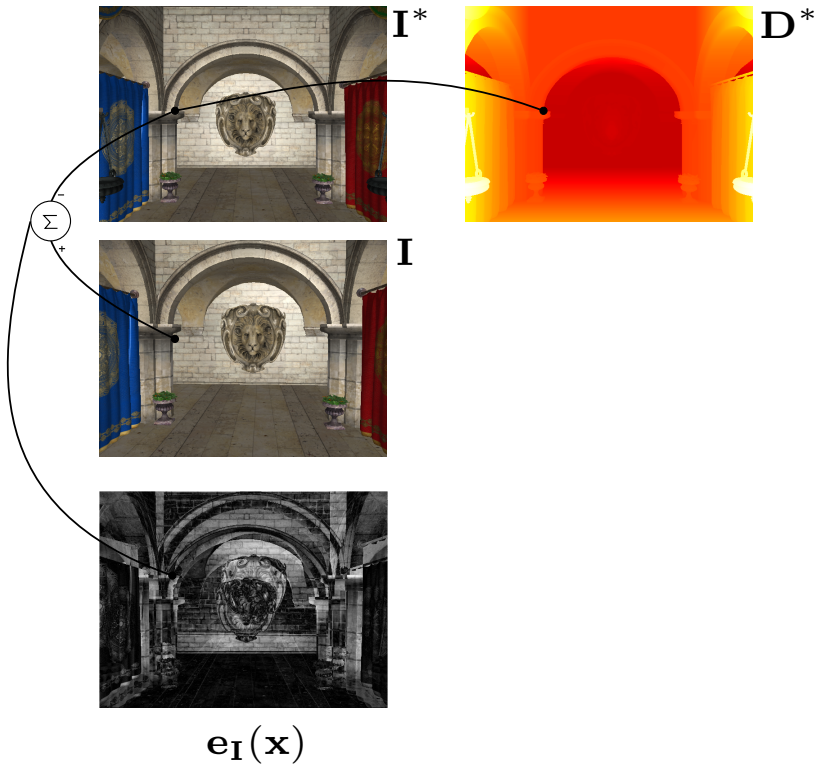
Sparse non-linear estimation

$$\mathbf{e_F}(\mathbf{x}) = w(\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}), \mathbf{v}^*) - \mathbf{p}$$

$$\mathbf{x} = \arg\min_{\mathbf{x}} \mathbf{e_F}^T \mathbf{W_F} \mathbf{e_F}$$

Geometric image-based term

- ✓ Computational efficiency
- ✓ Convergence domain
- ✗ Feature extraction and matching errors



- $\mathbf{w}(\mathbf{x})$ is the image warping function which depends on the pose and 3D points
- $\mathbf{p}$ are the 2D image points

Direct Image-based pose estimation

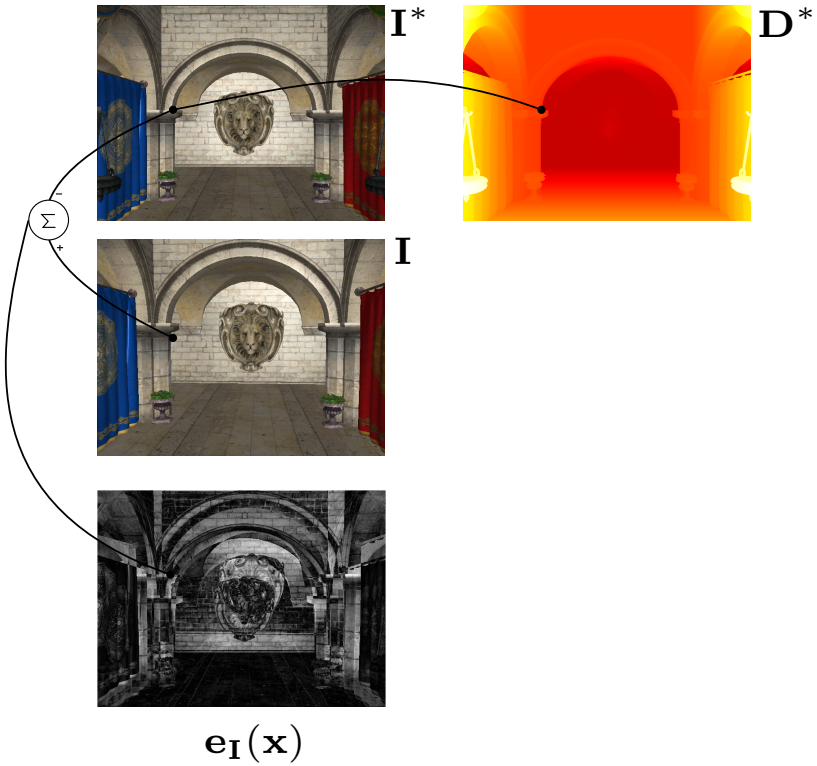
Direct color iterative registration

Efficient and accurate minimisation

$$\mathbf{e_I}(\mathbf{x}) = \mathbf{I}(w(\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}), \mathbf{v}^*)) - \mathbf{I}^*(\mathbf{p}^*)$$

$$\mathbf{x} = \arg\min_{\mathbf{x}} \mathbf{e}_\mathbf{I}^T \mathbf{W}_\mathbf{I} \mathbf{e}_\mathbf{I}$$

- $\mathbf{I}^*$ are the reference image intensities
- $\mathbf{I}(.)$ is the interpolation of the current image at $w(.)$



Matching and correspondance

Matching

- ✓ Sparse and efficient matching
- ✓ Matching performed globally across the image
- ✗ Requires feature extraction (can be costly and erroneous)
- ✗ Feature matching is local (can be costly and erroneous)
- ✗ Feature dependant - Harris, SIFT, SURF

Correspondance

- ✓ Efficient correspondances via interpoation
- ✓ Locally accurate (nearest neighbour bilinear, cubic,...)
- ✓ Uses global knowledge about the pose
- ✗ Only local - no global matches are possible

Matching and correspondance

Matching

- ✓ Sparse and efficient matching
- ✓ Matching performed globally across the image
- ✗ Requires feature extraction (can be costly and erroneous)
- ✗ Feature matching is local (can be costly and erroneous)
- ✗ Feature dependant - Harris, SIFT, SURF

Correspondance

- ✓ Efficient correspondances via interpoation
- ✓ Locally accurate (nearest neighbour bilinear, cubic,...)
- ✓ Uses global knowledge about the pose
- ✗ Only local - no global matches are possible

Matching and correspondance

Matching

- ✓ Sparse and efficient matching
- ✓ Matching performed globally across the image
- ✗ Requires feature extraction (can be costly and erroneous)
- ✗ Feature matching is local (can be costly and erroneous)
- ✗ Feature dependant - Harris, SIFT, SURF

Correspondance

- ✓ Efficient correspondances via interpoation
- ✓ Locally accurate (nearest neighbour bilinear, cubic,...)
- ✓ Uses global knowledge about the pose
- ✗ Only local - no global matches are possible

Feature-based vs Direct

Feature-based - sparse approaches

- ✓ Converges without texture overlap,
- ✓ Fast and efficient with minimal numbers of features,
- ✓ Redundancy between frames easily handled,
- ✓ Standard geometric tools can be used,
- ✗ Requires feature extraction (can be costly and erroneous)
- ✗ Reqires feature matching (can be costly and erroneous)
- ✗ Less robust than dense approaches
- ✗ Feature Extraction and Matching does not necessairily exploit global pose (Linearisation of the structure geometry)

Direct - dense approaches

- ✓ Models occlusions, free space
- ✓ Easy mesh extraction and path planning
- ✓ Super resolution possible in real-time Meilland and Comport (2013b)
- ✓ Sensor based -> accurate and robust
- ✓ Efficient and consistent loop closures/BA with standard tools
- ✓ Models geometric and photometric variation globally
- ✗ Redudancy between key-frames not easily handled
- ✗ Local convergence domain
- ✗ Dense and inefficient (also an advantage for applications requiring surfaces)

Feature-based vs Direct

Feature-based - sparse approaches

- ✓ Converges without texture overlap,
- ✓ Fast and efficient with minimal numbers of features,
- ✓ Redundancy between frames easily handled,
- ✓ Standard geometric tools can be used,
- ✗ Requires feature extraction (can be costly and erroneous)
- ✗ Reqires feature matching (can be costly and erroneous)
- ✗ Less robust than dense approaches
- ✗ Feature Extraction and Matching does not necessairily exploit global pose (Linearisation of the structure geometry)

Direct - dense approaches

- ✓ Models occlusions, free space
- ✓ Easy mesh extraction and path planning
- ✓ Super resolution possible in real-time Meilland and Comport (2013b)
- ✓ Sensor based -> accurate and robust
- ✓ Efficient and consistent loop closures/BA with standard tools
- ✓ Models geometric and photometric variation globally
- ✗ Redudancy between key-frames not easily handled
- ✗ Local convergence domain
- ✗ Dense and inefficient (also an advantage for applications requiring surfaces)

Feature-based vs Direct

Feature-based - sparse approaches

- ✓ Converges without texture overlap,
- ✓ Fast and efficient with minimal numbers of features,
- ✓ Redundancy between frames easily handled,
- ✓ Standard geometric tools can be used,
- ✗ Requires feature extraction (can be costly and erroneous)
- ✗ Reqires feature matching (can be costly and erroneous)
- ✗ Less robust than dense approaches
- ✗ Feature Extraction and Matching does not necessairily exploit global pose (Linearisation of the structure geometry)

Direct - dense approaches

- ✓ Models occlusions, free space
- ✓ Easy mesh extraction and path planning
- ✓ Super resolution possible in real-time Meilland and Comport (2013b)
- ✓ Sensor based -> accurate and robust
- ✓ Efficient and consistent loop closures/BA with standard tools
- ✓ Models geometric and photometric variation globally
- ✗ Redudancy between key-frames not easily handled
- ✗ Local convergence domain
- ✗ Dense and inefficient (also an advantage for applications requiring surfaces)

New unified direct and feature-based color pose estimation

Image-based registration with matching

Redefine optimisation criteria as:

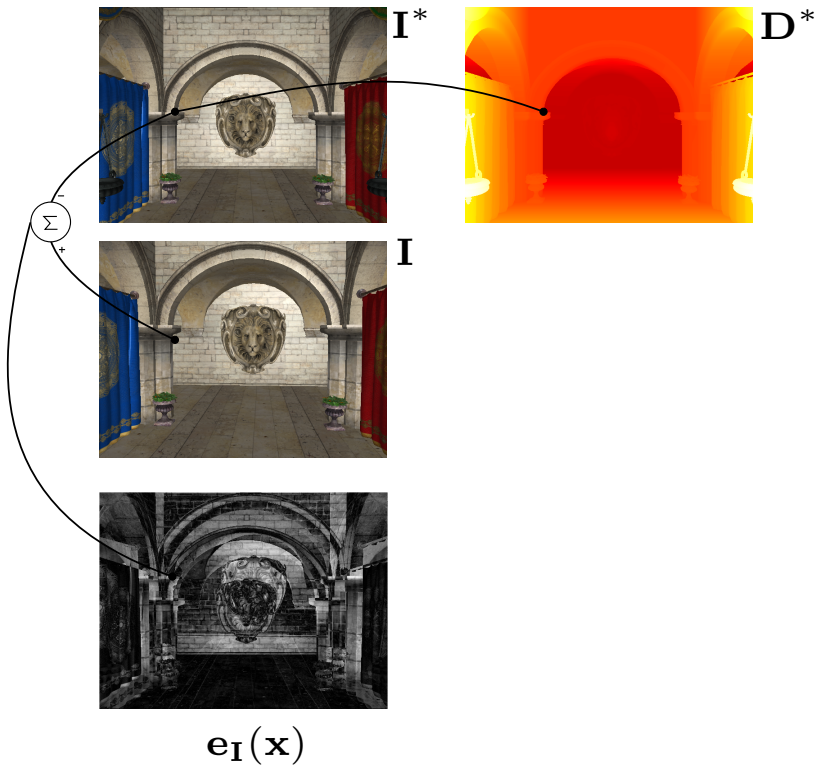
Define the measurements as
$\mathbf{M}_i = (u_i, v_i, I_i)$,

$$\mathbf{e_M}(\mathbf{x}) = \mathbf{M}(w(\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}), \mathbf{v}^*)) - \mathbf{M}^*(\mathbf{p}^*)$$

$$\mathbf{x} = \arg\min_{\mathbf{x}} \mathbf{e}_\mathbf{I}^T \mathbf{W}_\mathbf{I} \mathbf{e}_\mathbf{I}$$

Unified error term

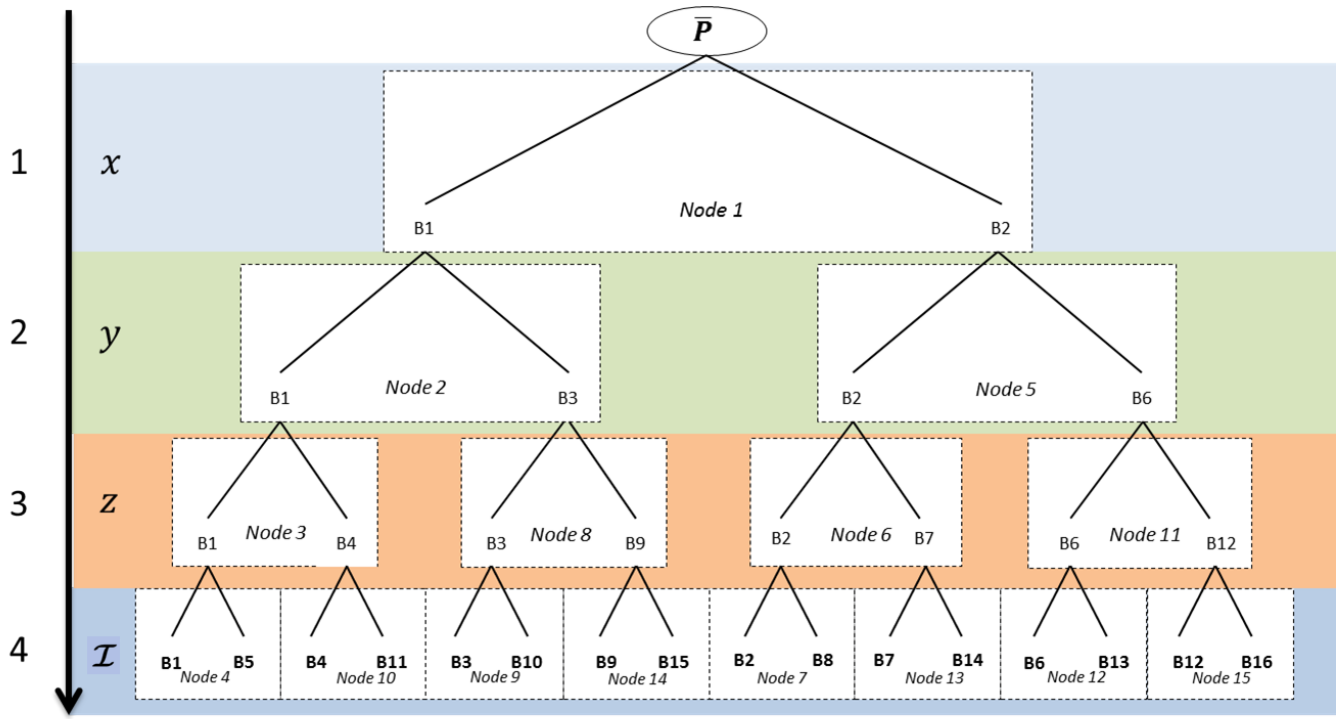
- ✓ Pure Direct approach when no matching
- ✓ Pure Feature-based approach when match gives the same color
- ✓ Improved convergence domain and efficiency



Efficient Direct Matching (no features)

Kd-tree

- Computed once for each reference image
- Nearest neighbour in color and 3D space
- Match perfomed only once on the first minimisation iteration



Kd-tree

Hybrid color and depth pose estimation

Dense color and depth iterative registration

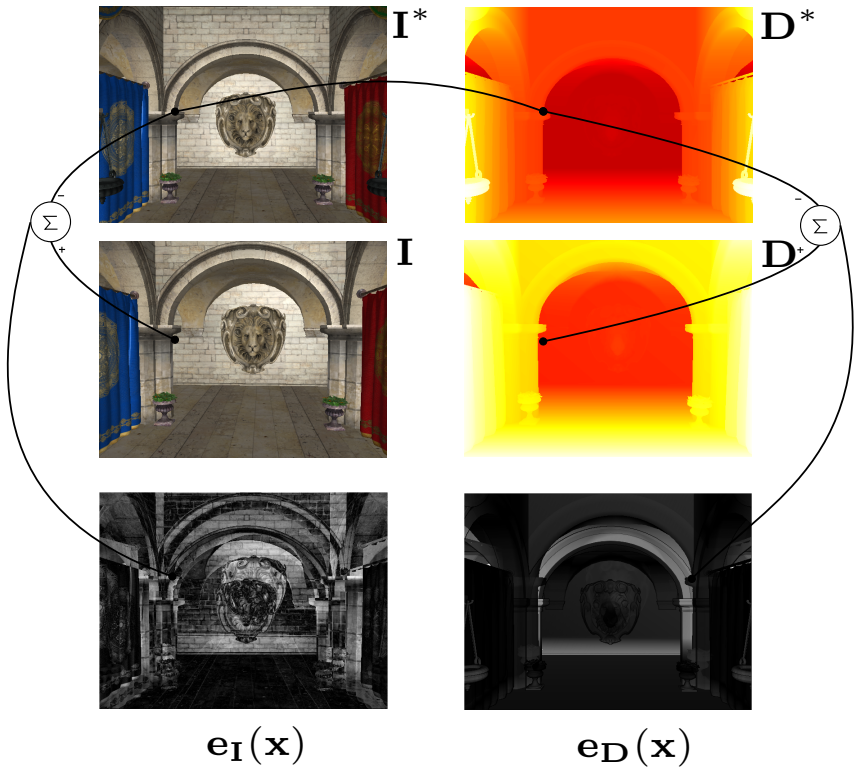
Efficient and robust bi-objective minimisation

$$\begin{cases} \mathbf{e_I}(\mathbf{x}) = \mathbf{I}(w(\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}), \mathbf{v}^*)) - \mathbf{I}^*(\mathbf{p}^*) \\ \mathbf{e_D}(\mathbf{x}) = \mathbf{N}(\mathbf{p})^T(\mathbf{v} - \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})\mathbf{v}^*) \end{cases}$$

$$\mathbf{x} = \arg\min_{\mathbf{x}} \lambda \mathbf{e_I}^T \mathbf{W_I} \mathbf{e_I} + \mathbf{e_D}^T \mathbf{W_D} \mathbf{e_D}$$

Hybrid minimisation

- ✓ Improved constrains, accuracy and robustness
- ✓ Better convergence domain and efficiency

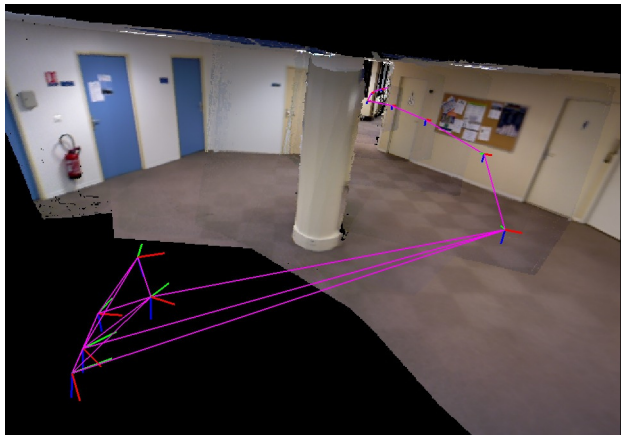


- λ can be chosen empirically or estimated simultaneously as in Kerl et al. (2013)

Model update / integration

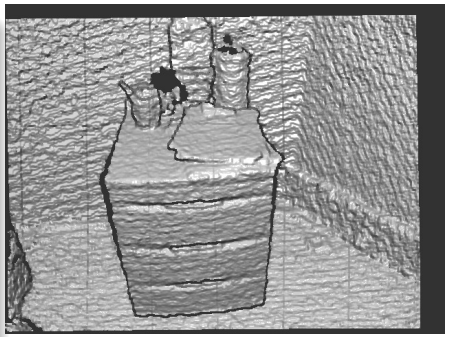
Graph management: new key-frame

- Added by monitoring the amount of occlusions/new information between the current frame and the next view prediction
- Connected topologically to the key-frames used for localisation

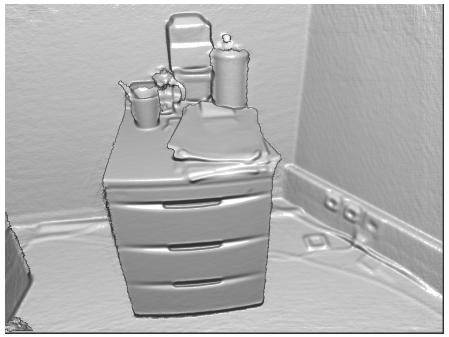


Local data fusion

- Local depth-maps improved by warping and fusing the current frame
- Super-resolution (*Meilland and Comport (2013b)*) ...



Original raw depth-map



Fused depth-map

Implementation

Full pipeline implemented on the GPU

- Algorithm highly parallel
- OpenGL
- Can also run on mobile GPU (Ultra-book)

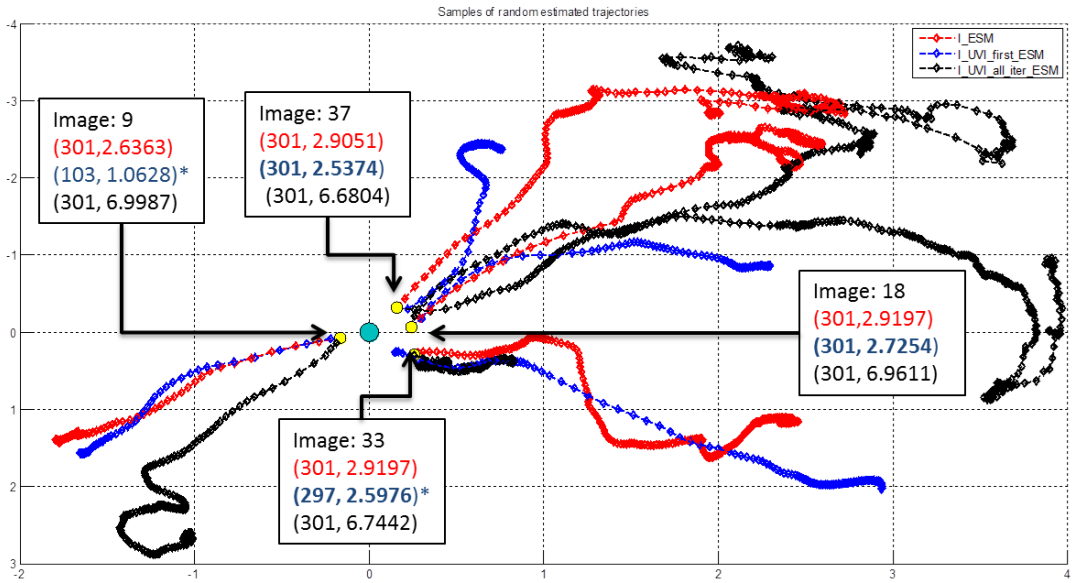
Setup: Nvidia GTX 670, Asus Xtion Pro Live, 640×480, 5 key frames for view prediction

Algorithm computing time

	Average time (ms)	Standard dev. (ms)
Initialisation	12.90	2.49
Pose estimation	11.78	2.06
Map update/integration	2.83	1.55
View Prediction	3.14	1.38
Total	30.63	3.08

Initialisation includes data transfers, image pyramids, bilateral filtering, Kd-tree ...

Quantitative results: Convergence and efficiency analysis from randomly generated poses



METHOD (Scale = 4)	Convergence rate	Number of iterations (average)	Time in sec (average)
Bilinear interpolation (I_ESM)	72	113.4200	1.0567
Alternative approach match at first iteration (I_UVI_first_ESM)	**80**	**99.3800**	**0.9232**
Alternative approach (I_UVI_all_iter_ESM)	54	182.2800	

Quantitative results: RGB-D datasets Sturm et al. (2012)

Methods

- P: Proposed approach,
- CKF: Closest key-frame
- Kintu.: Kintunuous (Whelan et al. 13)

Results

- Both image-based techniques outperform Kintunuous
  - Keeping raw images is crucial for pose accuracy.
- Proposed performs better than CKF
  - Maximises overlapping
  - View prediction also reduces self occlusions

Sequence	P (m/s)	CKF (m/s)	Kintu. (m/s)
fr1/desk	**0.0259**	0.0721	0.0393
fr2/desk	**0.0147**	0.0161	0.0208
fr1/room	**0.0351**	0.0502	0.0622
fr2/large_no_loop	**0.0695**	0.1529	0.1795

Relative root mean square error of drift in meters/second (RMSE).

Sequence	P (proposed)		CKF		Kintu.	
	Median	Max	Median	Max	Median	Max
fr1/desk	**0.018**	**0.066**	0.044	0.131	0.069	0.234
fr2/desk	**0.093**	**0.116**	0.099	0.130	0.119	0.362
fr1/room	**0.144**	**0.339**	0.201	0.430	0.158	0.421
fr2/large_no_loop	**0.187**	**0.317**	0.228	0.437	0.256	0.878

Absolute trajectory error in meters (ATE).

Quantitative results: RGB-D datasets Sturm et al. (2012)

Methods

- P: Proposed approach,
- CKF: Closest key-frame
- Kintu.: Kintunuous (Whelan et al. 13)

Results

- Both image-based techniques outperform Kintunuous
  - Keeping raw images is crucial for pose accuracy.
- Proposed performs better than CKF
  - Maximises overlapping
  - View prediction also reduces self occlusions

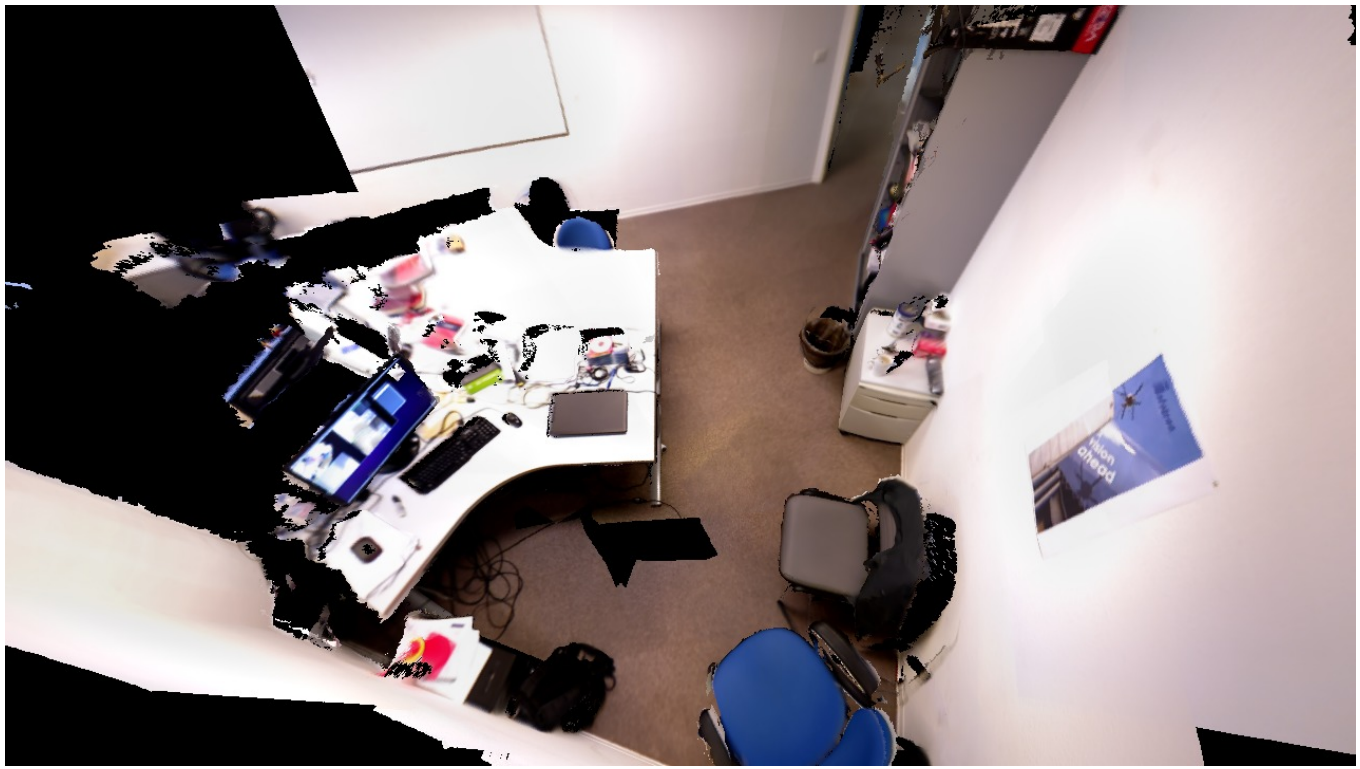
Sequence	P (m/s)	CKF (m/s)	Kintu. (m/s)
fr1/desk	**0.0259**	0.0721	0.0393
fr2/desk	**0.0147**	0.0161	0.0208
fr1/room	**0.0351**	0.0502	0.0622
fr2/large_no_loop	**0.0695**	0.1529	0.1795

Relative root mean square error of drift in meters/second (RMSE).

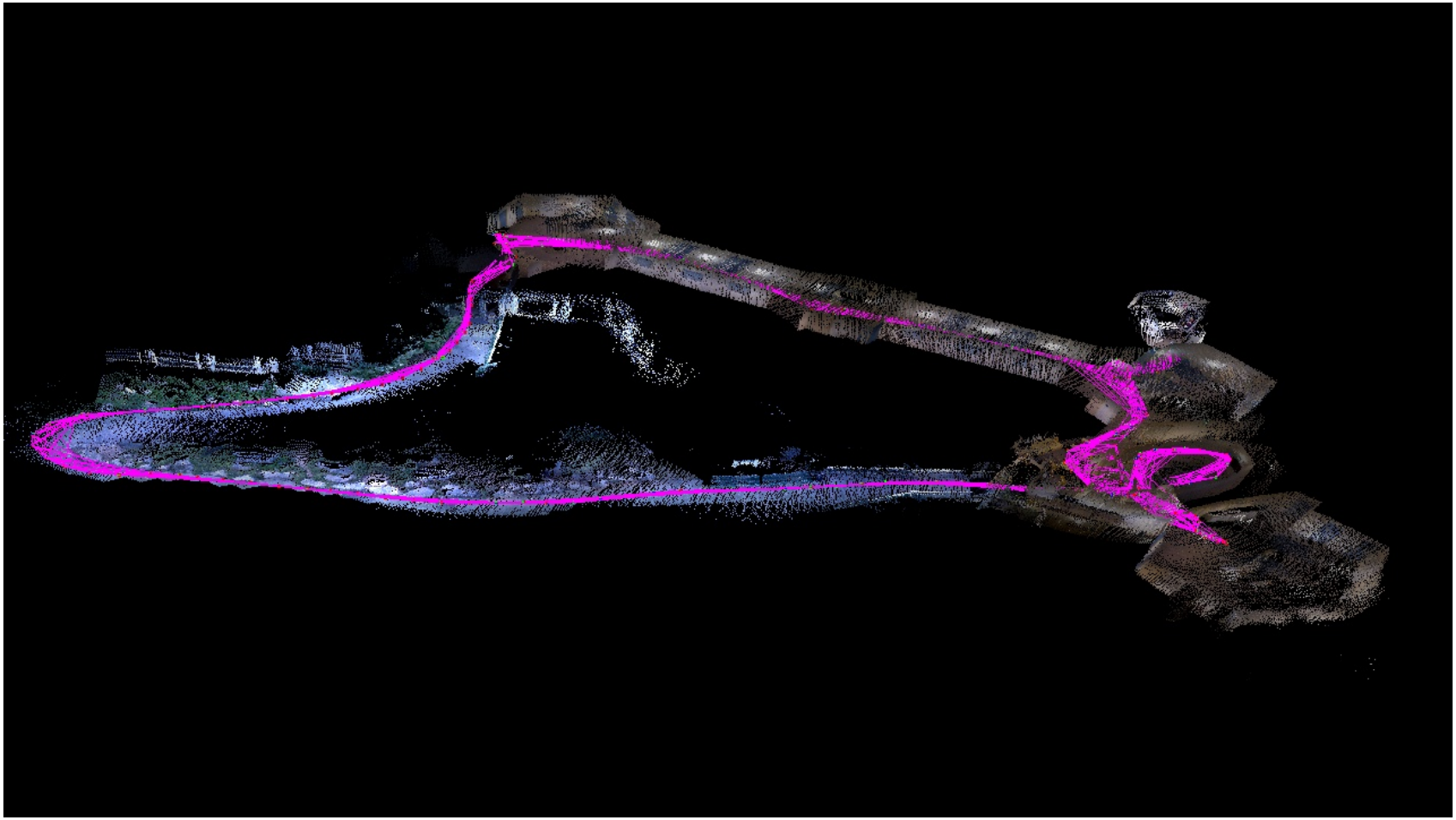
Sequence	P (proposed)		CKF		Kintu.	
	Median	Max	Median	Max	Median	Max
fr1/desk	**0.018**	**0.066**	0.044	0.131	0.069	0.234
fr2/desk	**0.093**	**0.116**	0.099	0.130	0.119	0.362
fr1/room	**0.144**	**0.339**	0.201	0.430	0.158	0.421
fr2/large_no_loop	**0.187**	**0.317**	0.228	0.437	0.256	0.878

Absolute trajectory error in meters (ATE).

Qualitative results: Office size



Recent developments, HDR, LBA, Loop Closure ...



High Dynamic Range mapping ISMAR 2013, Rolling shutter and motion blur ICCV 2013

Conclusion

Real-time dense localisation and mapping approach

Efficient key-frame representation

- Allows to map **large scale** environments
- Accuracy of raw images

Dense bi-objective minimisation

- Combines **photometric** and **depth** errors terms
- Robust and accurate localisation

New direct matching approach

- Improves **convergence domain** and **efficiency**

Conclusion

Real-time dense localisation and mapping approach

Efficient key-frame representation

- Allows to map **large scale** environments
- Accuracy of raw images

Dense bi-objective minimisation

- Combines **photometric** and **depth** errors terms
- Robust and accurate localisation

New direct matching approach

- Improves **convergence domain** and **efficiency**

Conclusion

Real-time dense localisation and mapping approach

Efficient key-frame representation

- Allows to map **large scale** environments
- Accuracy of raw images

Dense bi-objective minimisation

- Combines **photometric** and **depth** errors terms
- Robust and accurate localisation

New direct matching approach

- Improves **convergence domain** and **efficiency**

Tendances in new sensor modalities

Tendancies in sensors

RGB-D sensors

- Smaller and more mobile sensors - i.e. Intel realsense
- Stereo embeded RDB-D sensors for outdoor use
- Laser combined with vision,...

Emerging sensors

- Dynamic Vision Sensors (DVS) - Event sesnors for high-speed SLAM - i.e.
- Light field sensors - i.e. Raytrix, Lytro,..

R5 light field camera

Visual Servoing and SLAM on the HRP2

Thank you for your attention.

References I

J. Chen, D. Bautembach, and S. Izadi. Scalable real-time volumetric surface reconstruction. *ACM Trans. Graph.*, 32(4), 2013.

A. I. Comport, E. Malis, and P. Rives. Accurate quadrifocal tracking for robust 3d visual odometry. In *IEEE Int. Conf. on Robotics and Automation*, 2007.

P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In *Int. Symp. on Experimental Robotics*, 2010.

C. Kerl, J. Sturm, and D. Cremers. Robust odometry estimation for RGB-D cameras. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2013.

G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. In *IEEE and ACM Int. Symp. on Mixed and Augmented Reality*, pages 1–10, 2007.

K. Konolige, M. Agrawal, R. Bolles, C. Cowan, M.Fischler, and B. Gerkey. Outdoor mapping and navigation using stereo vision. In *Intl. Symp. on Experimental Robotics (ISER)*, 2007.

M. Meilland and A. Comport. On unifying key-frame and voxel-based dense visual SLAM at large scales. In *International Conference on Intelligent Robots and Systems*, Tokyo, Japan, 3-8 November 2013a. IEEE/RSJ. URL `http://www.i3s.unice.fr/~comport/publications/2013_IROS_Meilland.pdf`.

M. Meilland and A. I. Comport. Super-resolution 3d tracking and mapping. In *IEEE Int. Conf. on Robotics and Automation*, 2013b.

M. Meilland, A. I. Comport, and P. Rives. A spherical robot-centered representation for urban navigation. In *IEEE Int. Conf. on Intelligent Robots and Systems*, 2010.

P. Mordohai, J. m. Frahm, A. Akbarzadeh, C. Engels, D. Gallup, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, H. Towles, G. Welch, R. Yang, M. Pollefeys, and D. Nistér. Real-time video-based reconstruction of urban environments. In *Proceedings of 3DARCH: 3D Virtual Reconstruction and Visualization of Complex Architectures*, 2007.

R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Int. symposium on mixed and augmented reality*, 2011.

J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Int. Conf. on Intelligent Robot Systems*, 2012.

T. Whelan, H. Johannsson, M. Kaess, J. Leonard, and J. McDonald. Robust real-time visual odometry for dense RGB-D mapping. In *IEEE Int. Conf. on Robotics and Automation*, 2013.