Integrating Algorithmic Parameters into Benchmarking and Design Space Exploration in 3D Scene Understanding

Luigi Nardi, PhD

Software Performance Optimisation group

@PACT Haifa

September 12th 2016

In collaboration with:

B. Bodin, M Z. Zia, Harry Wagstaff, G. S. Shenoy, M. K. Emani, J. Mawer, C. Kotselidis, A. Nisbet, M. Luján, B. Franke, P. H. J. Kelly and M. F. P. O'Boyle

Take away messages

1. Our community has to move from benchmarks that are used like unit tests, towards benchmarks that expose performance/energy tradeoffs against quantitative end-to-end quality of result.

2. Doing so opens up the scope for adaptivity, and helps ensure we are optimising at lower levels for the right profile of higher-level operations.

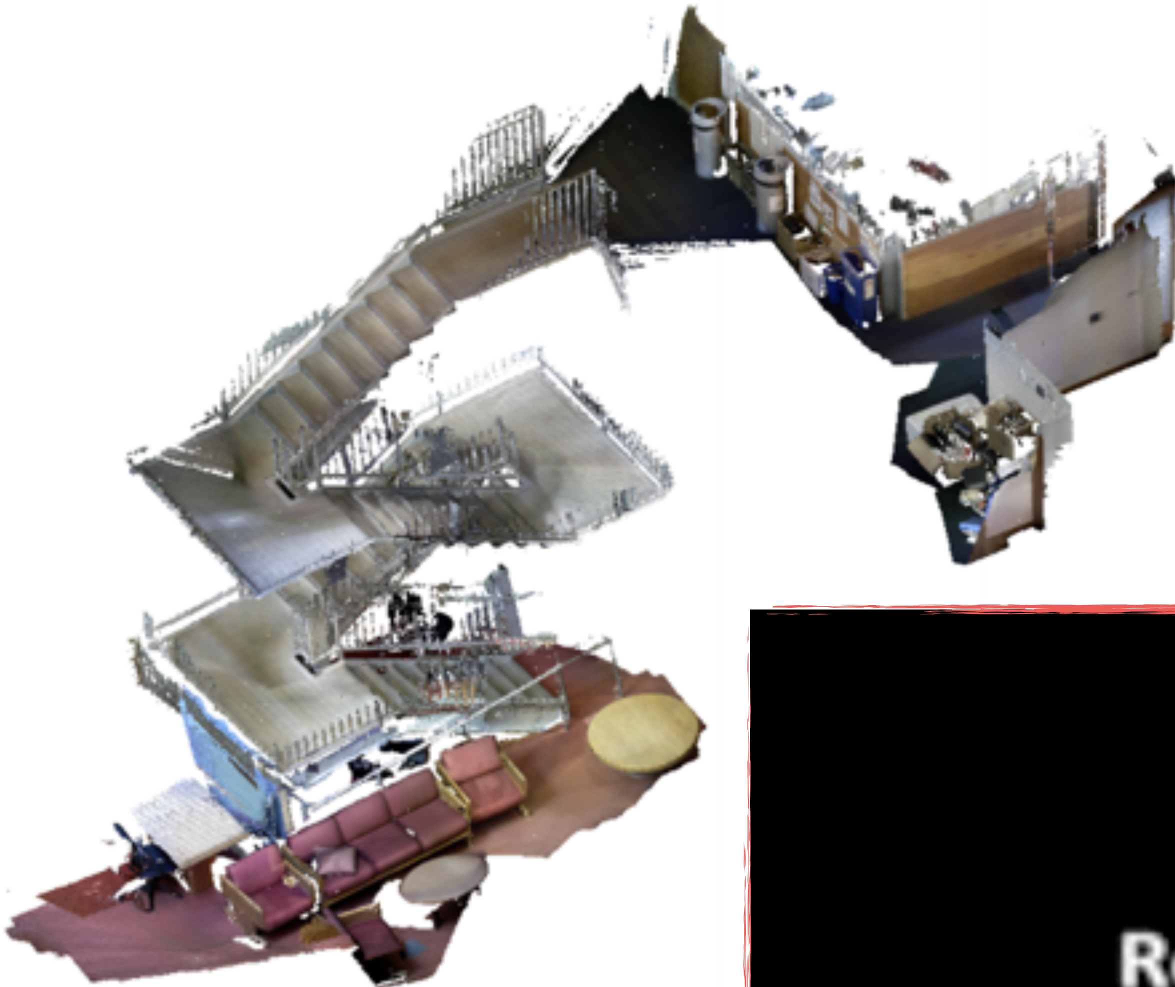
3. Some of the techniques we use for auto-tuning at the compiler/architecture level can be used at the algorithmic adaptation level - in fact achieving much bigger performance improvements.

Simultaneous localisation and mapping (SLAM)

Build a coherent world representation and localise the camera in real-time

Applications, e.g.:

- Robotics
- Autonomous driving
- 3D printing
- Augmented reality



[Whelan et al. 2012]

Video:
[Newcombe et al. ISMAR 2011]

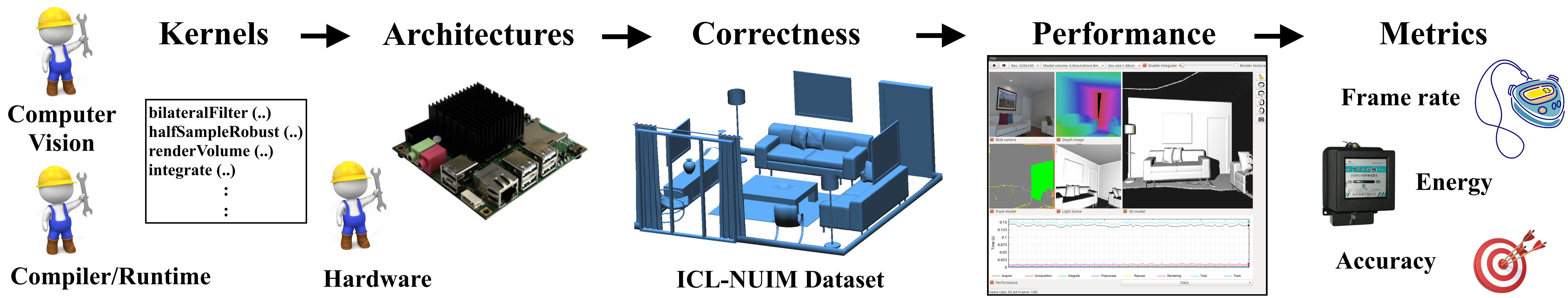
What CV researchers say about KinectFusion performance

"KinectFusion cannot run in real-time on mobile"

"You need a fat GPU to run KFusion"

Holistic approach to SLAM performance:

SLAMBench

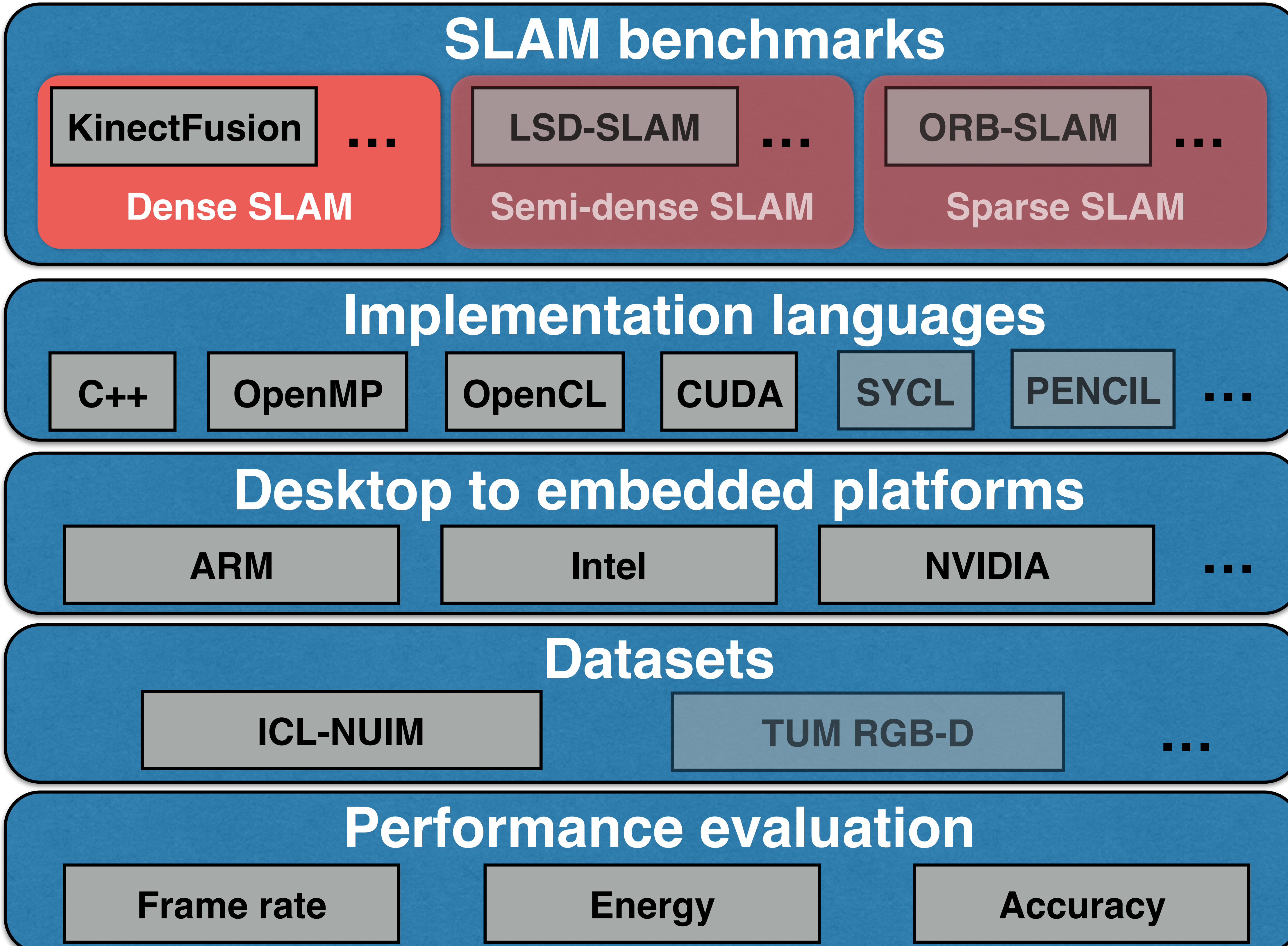


A publicly-available benchmarking framework for quantitative, comparable and validatable experimental research to investigate trade-offs in performance, accuracy and energy consumption of a SLAM system

Error metric: absolute trajectory error (ATE) based on dataset ground truth

Introducing SLAMBench, a performance and accuracy benchmarking methodology for SLAM **(ICRA 2015)**

SLAMBench framework



What is the optimisation space?

Configuration parameters:

Co-design space

Space 1

1. Algorithmic:
   - Application-specific parameters
   - Minimisation methods
   - Early exit condition values

Space 2

2. Compilation:
   - opencl-params: -cl-mad-enable,-cl-fast-relaxed-math, etc.
   - LLVM flags: O1, O2, O3, vectorize-slp-aggressive, etc.
   - Local work group size: 16/32/64/96/112/128/256
   - Vectorisation: width (1/2/4/8), direction (x/y)
   - Thread coarsening: factor (1/2/4/8/16/32), stride (1/2/4/8/16/32), dimension (x/y)

Space 3

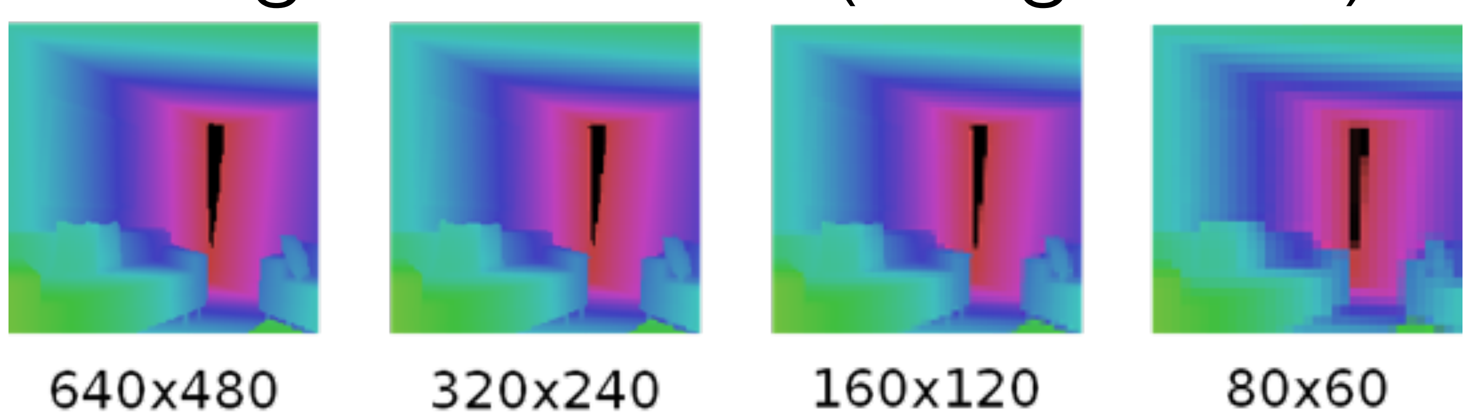
3. Architecture:
   - GPU frequency: 177/266/350/420/480/543/600/DVFS
   - # of active big cores: 0/1/2/3/4
   - # of active LITTLE cores: 1/2/3/4

Warning: huge spaces, impossible to run exhaustively

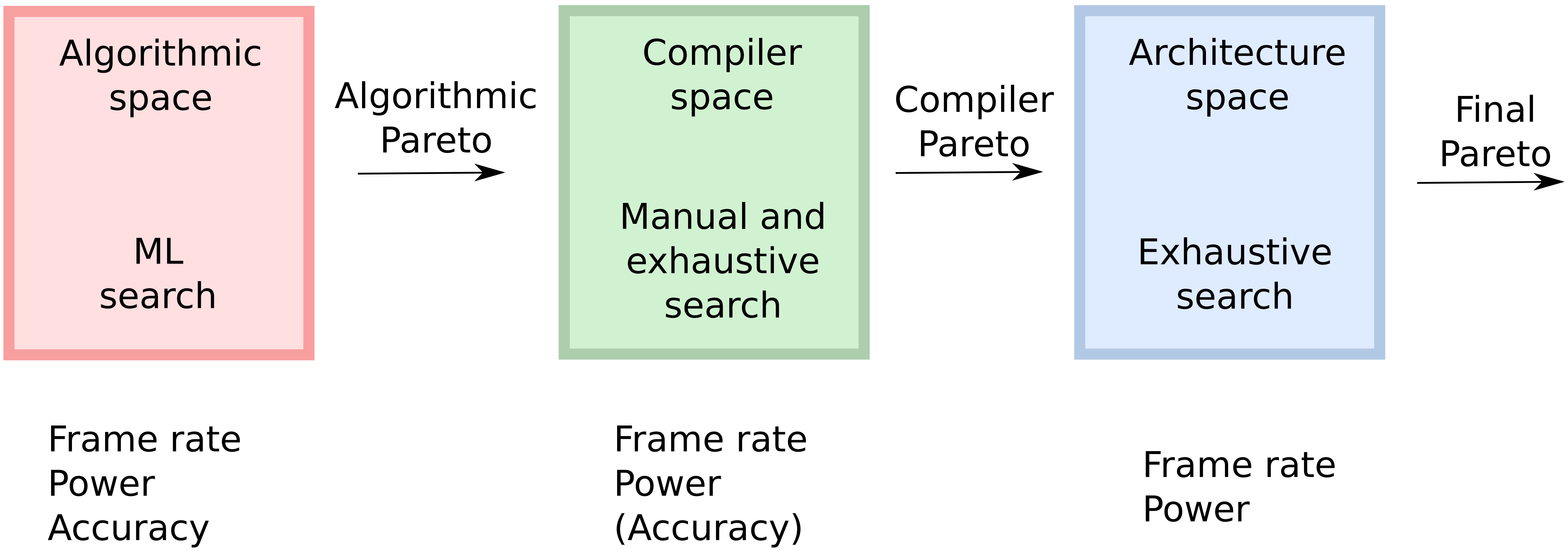
KinectFusion algorithmic features

Features	Ranges
Volume resolution	64x64x64, 128x128x128, 256x256x256, 512x512x512
μ distance	0 .. 0.5
Pyramid level iterations (3 levels)	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
Image resolution (image ratio)	1, 2, 4, 8
Tracking rate	1, 2, 3, 4, 5
ICP threshold	10^{-6} .. 10^{2}
Integration rate	1 .. 30

Image resolution (image ratio)

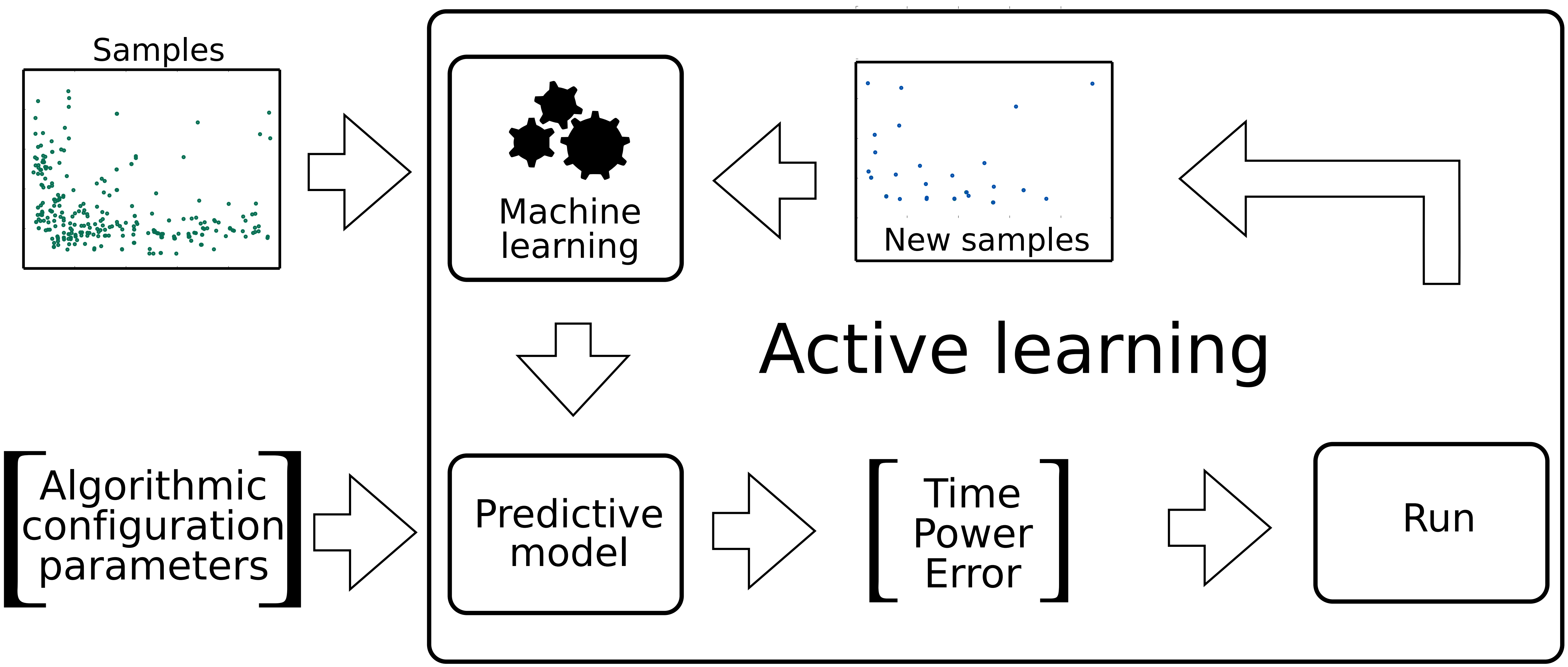


Incremental exploration approach



Incremental optimisation process is not the way to achieve an optimal result, would be better to explore the full product space

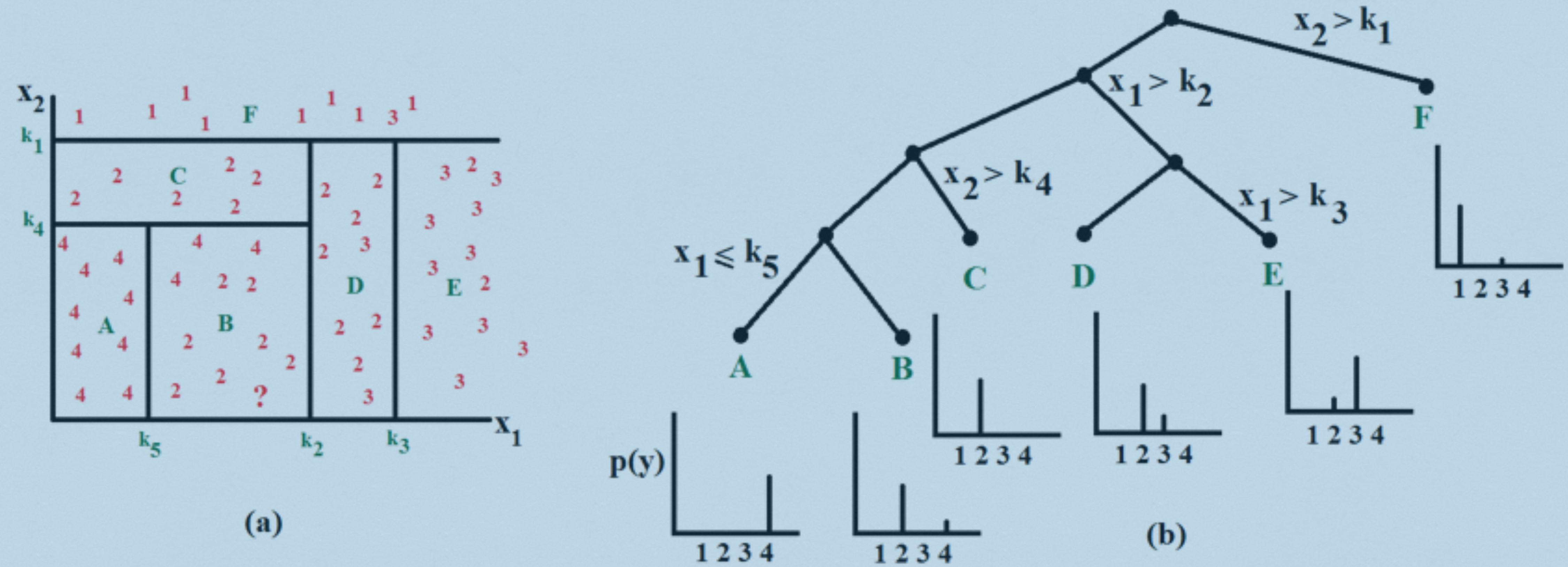
Algo design-space exploration (DSE)



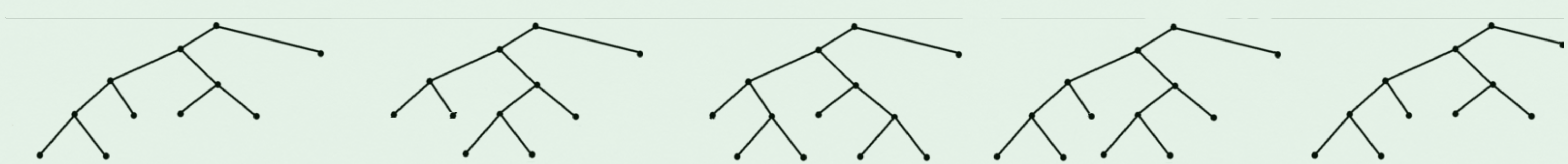
Luigi Nardi - Imperial College London

Machine learning methods used

Decision Tree

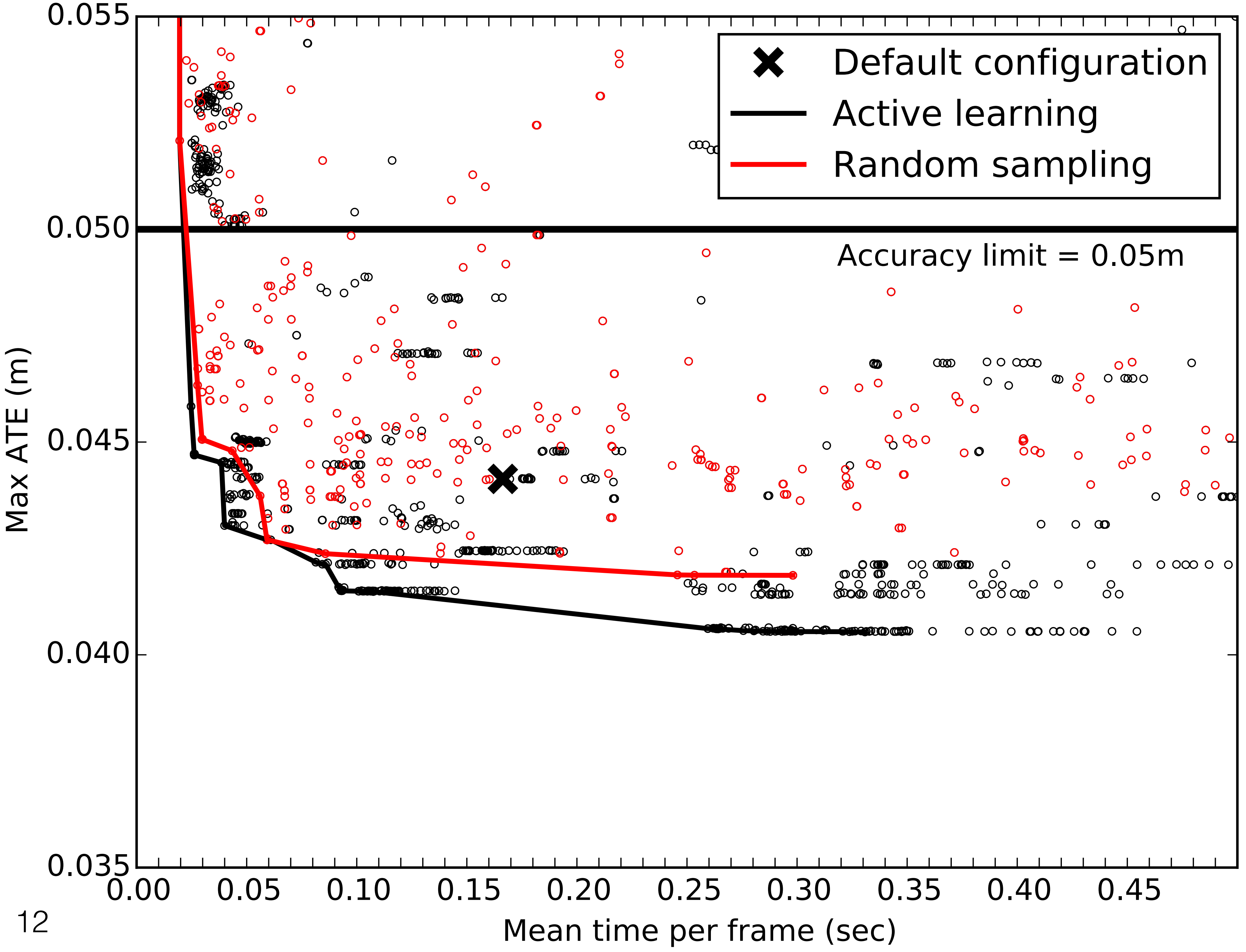


Random Forest

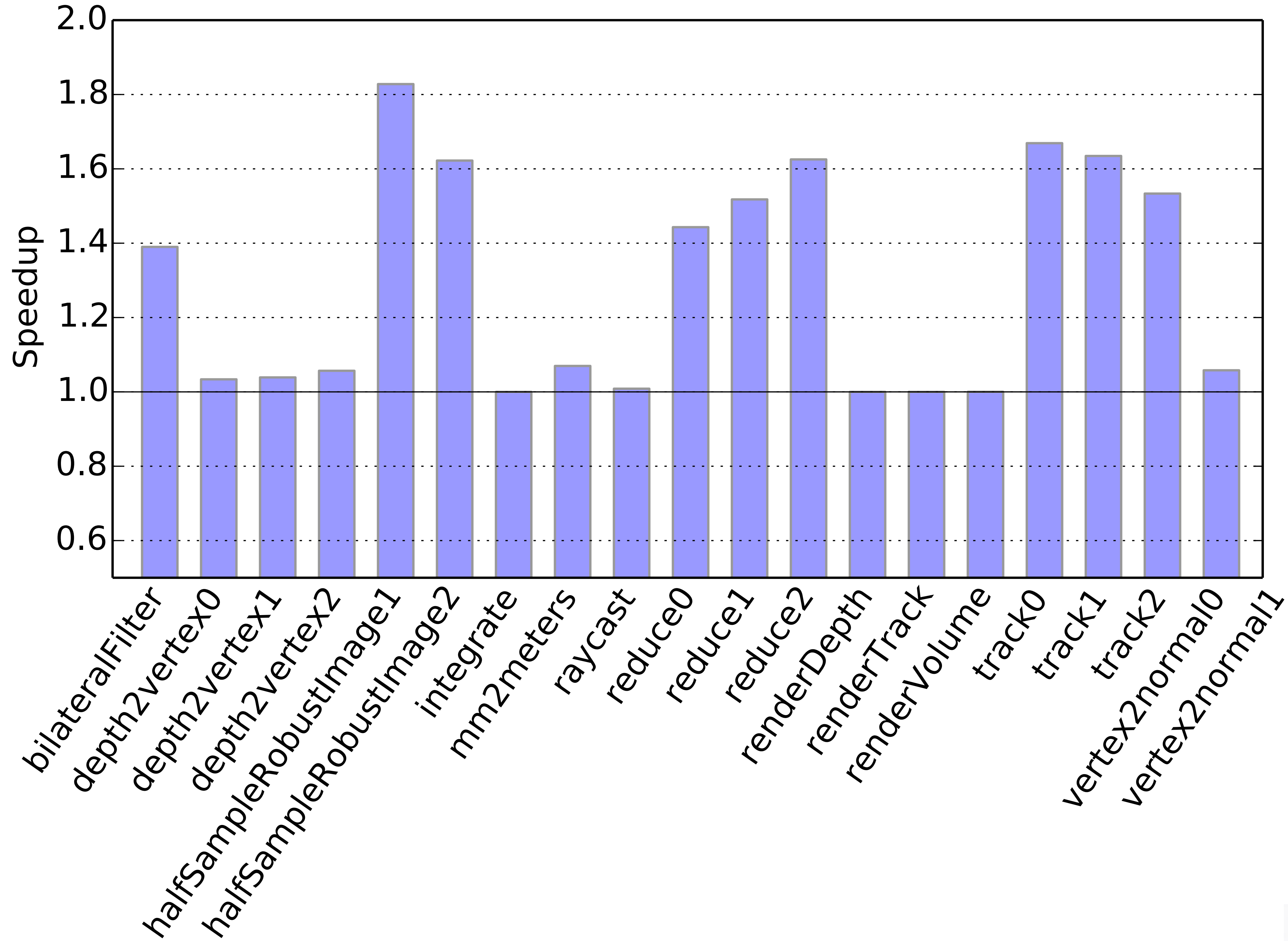


DSE on algorithmic parameters error/runtime

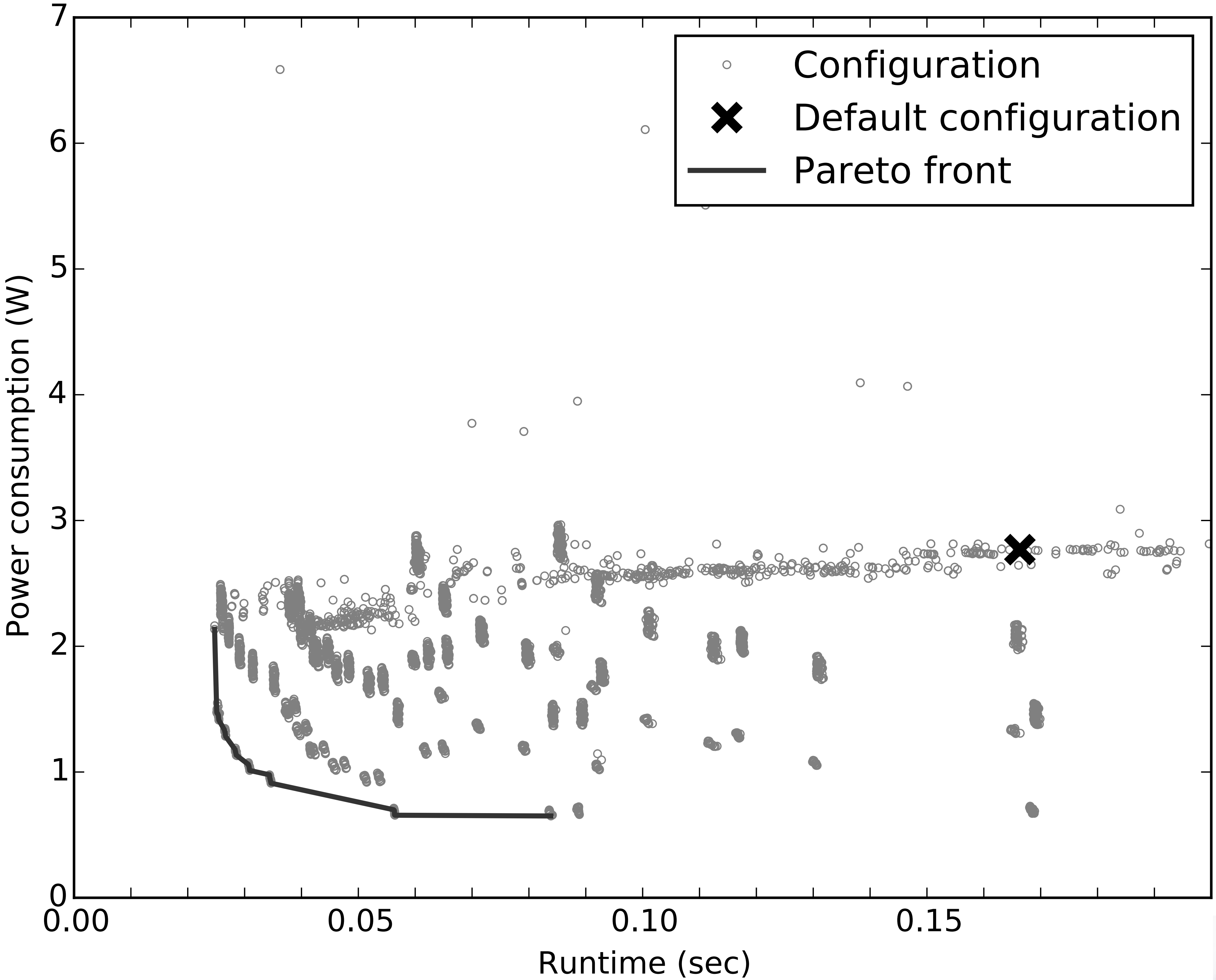
Machine	CPU	CPU name	CPU GFLOPS	CPU cores	GPU	GPU name	GPU GFLOPS	TDP Watts
Hardkernel **ODROID-XU3**	ARM A15 + A7	Exynos 5422	80	4 + 4	ARM	Mali-T628	60 + 30	10



DSE compiler parameters speedup



DSE architecture parameters power/runtime

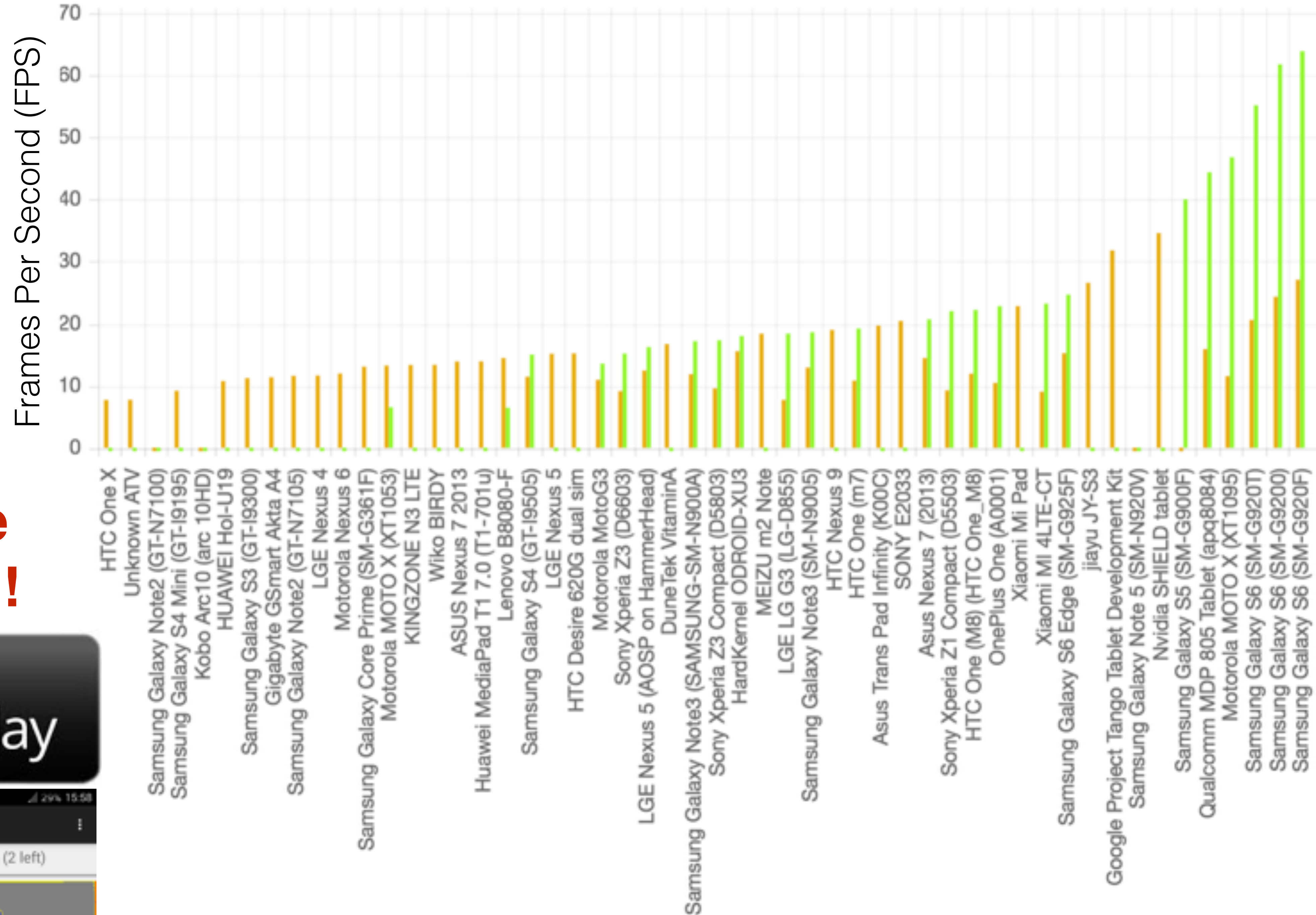


DSE final result

Constraint	Runtime (FPS)	Max ATE (cm)	Power (Watts)
Default	6.03	4.41	2.77
Best runtime	39.85	4.47	1.47
Best accuracy	1.51	3.30	2.38
Best power	11.92	4.45	0.65
Power < 1W	29.09	4.47	0.98
Power < 2W	39.85	4.47	1.47
FPS > 10	11.92	4.45	0.65
FPS > 20	28.87	4.47	0.91
FPS > 30	32.38	4.47	1.01

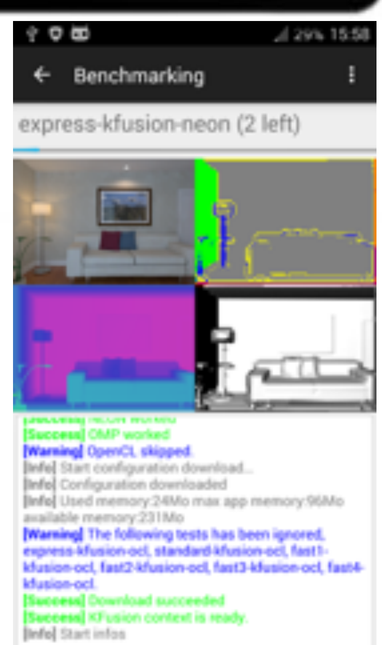
- Most of the improvement comes from the algorithmic space
- KinectFusion real-time on a popular embedded device
- Enabling auto-tuning at the domain-specific level

Crowdsourcing mobile Android SLAMBench



Get it now, and see where your device is!!

Get it on Google play



- It runs a set of Pareto configurations on the available languages on your device
- Then shows the best achieved result

Conclusion - take away messages

1. Our community has to move from benchmarks that are used like unit tests, towards benchmarks that expose performance/energy tradeoffs against quantitative end-to-end quality of result.

2. Doing so opens up the scope for adaptivity, and helps ensure we are optimising at lower levels for the right profile of higher-level operations.

3. Some of the techniques we use for auto-tuning at the compiler/architecture level can be used at the algorithmic adaptation level - in fact achieving much bigger performance improvements.

References I

- [Nardi et al. 2015] L. Nardi, B. Bodin, M. Z. Zia, J. Mawer, A. Nisbet, P. H. J. Kelly, A. J. Davison, M. Luján, M. F. P. O'Boyle, G. Riley, N. Topham, and S. Furber. "Introducing SLAMBench, a performance and accuracy benchmarking methodology for SLAM." Submitted, arXiv:1410.2167, 2015.
- [Newcombe et al. ICCV 2011] R. A. Newcombe, S. J. Lovegrove and A. J. Davison. "DTAM: Dense tracking and mapping in real-time." Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011.
- [Rusinkiewicz and Levoy 2001] S. Rusinkiewicz, and M. Levoy. "Efficient variants of the ICP algorithm." 3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on. IEEE, 2001.
- [Chen et al. 2013] J. Chen, D. Bautembach, and S. Izadi, Scalable real-time volumetric surface reconstruction, in ACM Trans. Graph., 2013.
- [Newcombe et al. ISMAR 2011] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. "KinectFusion: Real-time dense surface mapping and tracking." 10th IEEE Int. Symp. on Mixed and augmented reality (ISMAR), 2011.
- [Handa et al. 2014] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A Benchmark for RGB-D Visual Odometry, 3D Reconstruction and SLAM. IEEE Int. Conf. on Robotics and Automation, ICRA 2014.
- [Reitmayr] G. Reitmayr. KFusion github 2011. https://github.com/GerhardR/kfusion
- [Curless and Levoy 1996] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In Proc. Computer graphics and interactive technique. ACM, 1996.
- [Whelan et al. 2012] T. Whelan, M. Kaess, M. Fallon,H. Johannsson, J. Leonard, and J. McDonald. Kintinuous: Spatially extended kinectfusion. 2012.
- C. Jiawen, D. Bautembach, and S. Izadi. "Scalable real-time volumetric surface reconstruction." ACM TOG, 2013.
- Frahm, Jan-Michael, et al. "Building Rome on a cloudless day." Computer Vision–ECCV 2010. Springer Berlin Heidelberg, 2010.
- Erhan, Dumitru, et al. "Scalable object detection using deep neural networks." Proceedings of the IEEE CVPR. 2014.

References II

- Arbelaez, Pablo, et al. "Contour detection and hierarchical image segmentation." IEEE Pattern Analysis and Machine Intelligence, 2011.
- [Ogilvie 2014] Ogilvie, William, et al. "Fast automatic heuristic construction using active learning." Proceedings of the Workshop on Languages and Compilers for Parallel Computing (LCPC'14). 2014.
- [Siegmund 2015] Siegmund Norbert et al. "Performance-influence models for highly configurable systems", submitted FSE 2015.
- [Guo 2013] Guo, Jianmei, et al. "Variability-aware performance prediction: A statistical learning approach." Automated Software Engineering (ASE), 2013 IEEE/ACM 28th International Conference on. IEEE, 2013.
- [Grewe 2011] Grewe, Dominik et al. "A static task partitioning approach for heterogeneous systems using OpenCL." Compiler Construction. Springer Berlin Heidelberg, 2011.
- [Kurek 2013] Kurek, Maciej, Tianchi Liu, and Wayne Luk. "MULTI-OBJECTIVE SELF-OPTIMIZATION OF RECONFIGURABLE DESIGNS WITH MACHINE LEARNING." 2nd Workshop on Self-Awareness in Reconfigurable Computing Systems (SRCS'13). 2013.
- [Balaprakash 2013] Balaprakash, Prasanna, Robert B. Gramacy, and Stefan M. Wild. "Active-learning-based surrogate models for empirical performance tuning." Cluster Computing (CLUSTER), 2013 IEEE International Conference on. IEEE, 2013.
- [Vespa 2015] Vespa Emanuele. "Sparse voxelization of dense volumetric reconstruction with automated analysis of scene reconstruction quality." M.Res. thesis, Imperial College London, 2015.

Copyrights

- Author: unknown. Microsoft Kinect camera. [Image]. Retrieved from http://channel9.msdn.com/Series/KinectSDKQuickstarts/Understanding-Kinect-Hardware
- Author: Dyson Ltd. Dyson 360 Eye. [Video]. Retrieved from https://www.youtube.com/watch?v=OadhuICDAjk
- Author: Google Inc. Google Tango project. [Image]. Retrieved from http://blogthinkbig.com/en/project-tango-googles-mobile-kinect/
- Author: unknown. Audi autonomous car. [Photograph]. Retrieved from http://www.wired.com/2010/06/audis-robotic-car-looks-hot-in-old-school-livery/
- Author: ExtremeTech. Google Shaft robot. [Photograph]. Retrieved from http://www.extremetech.com/extreme/173318-google-wins-darpas-robotics-challenge-wonders-if-it-was-a-good-idea-to-turn-down-future-military-contracts
- Author: HardKernel. ODROID-XU3 board. [Photograph]. Retrieved from http://www.hardkernel.com/main/products/prdt_info.php?g_code=G135235611947
- Author: PC Specialist Ltd. Vortex series laptop. [Photograph]. Retrieved from https://www.pcspecialist.co.uk/forums/showthread.php?23366-My-new-beast-15-6-quot-Vortex-III
- Author: Arndale.org. Arndale board. [Photograph]. Retrieved from http://www.arndaleboard.org/wiki/index.php/Main_Page
- Author: Unknown. Chip. [Image]. Retrieved from https://cajalesygalileos.wordpress.com/2013/06/23/un-chip-ultrasensible-identifica-15-cepas-de-gripe/
- Author: Unknown. Eye. [Image]. Retrieved from http://gallery.digitalculture.asu.edu/?/interactive-environments/computer-vision/
- Author: Unknown. Compiler. [Image]. Retrieved from http://d3q6qq2zt8nhwv.cloudfront.net/107/large-icon.png