Tombone's Computer Vision Blog

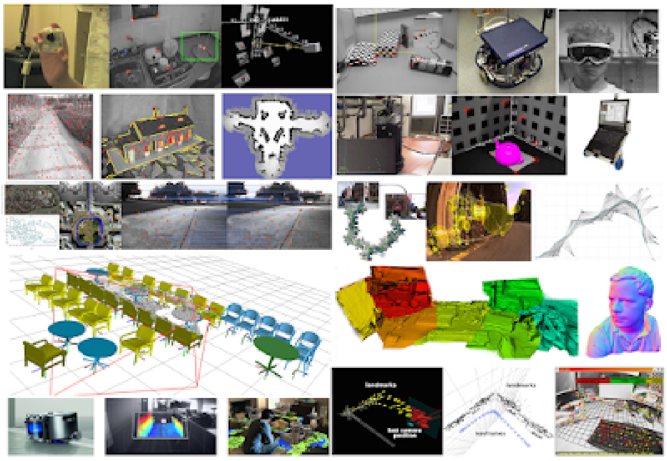
Deep Learning, Computer Vision, and the algorithms that are shaping the future of Artificial Intelligence.

Wednesday, January 13, 2016

The Future of Real-Time SLAM and Deep Learning vs SLAM

Last month's International Conference of Computer Vision (ICCV) was full of Deep Learning techniques, but before we declare an all-out ConvNet victory, let's see how the other "non-learning" geometric side of computer vision is doing. **S**imultaneous **L**ocalization **and** **M**apping, or **SLAM**, is arguably one of the most important algorithms in Robotics, with pioneering work done by both computer vision and robotics research communities. Today I'll be summarizing my key points from ICCV's Future of Real-Time SLAM Workshop, which was held on the last day of the conference (December 18th, 2015).

Today's post contains a brief introduction to SLAM, a detailed description of what happened at the workshop (with summaries of all 7 talks), and some take-home messages from the *Deep Learning-focused panel discussion* at the end of the session.



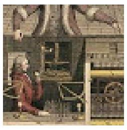
The Future of Real-Time SLAM: 18th December 2015 (ICCV Workshop)

SLAM visualizations. Can you identify any of these SLAM algorithms?

Part I: Why SLAM Matters

Visual SLAM algorithms are able to simultaneously build 3D maps of the world while tracking the location and orientation of the camera (hand-held or head-mounted for AR or mounted on a robot). SLAM algorithms are complementary to ConvNets and Deep Learning: SLAM focuses on geometric problems and Deep Learning is the master of perception (recognition) problems. If you want a robot to go towards your refrigerator without hitting a wall, use SLAM. If you want the robot to identify the items inside your fridge, use ConvNets.

Popular Posts

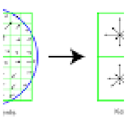


Deep Learning vs Machine Learning vs Pattern Recognition

Lets take a close look at three related terms (Deep Learning vs Machine Learning vs Pattern Recognition), and see how they relate to some o...

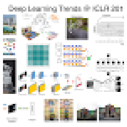
The Future of Real-Time SLAM and Deep Learning vs SLAM

Last month's International Conference of Computer Vision (ICCV) was full of Deep Learning techniques, but before we declare an all-out...



From feature descriptors to deep learning: 20 years of computer vision

We all know that deep convolutional neural networks have produced some stellar results on object detection and recognition benchmarks in th...



Deep Learning Trends @ ICLR 2016

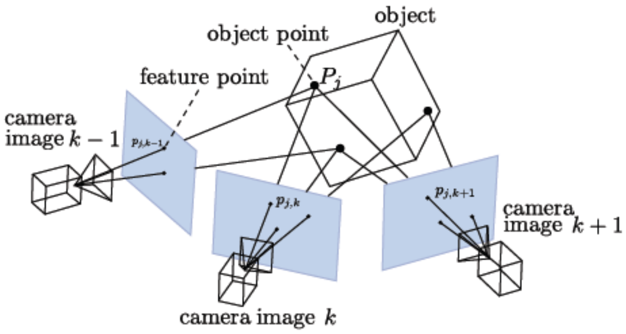
Started by the youngest members of the Deep Learning Mafia [1], namely Yann LeCun and Yoshua Bengio , the ICLR conference is quickly becom...

ICCV 2015: Twenty one hottest research papers

"Geometry vs Recognition" becomes ConvNet-for-X Computer Vision used to be cleanly separated into two schools: geometry and rec...

Deep Learning vs Probabilistic Graphical Models vs Logic

Today, let's take a look at three paradigms that have shaped the field of Artificial Intelligence in the last 50 years: Logic , Probab...



Basics of SfM/SLAM: From point observation and intrinsic camera parameters, the 3D structure of a scene is computed from the estimated motion of the camera. For details, see openMVG website.

SLAM is a real-time version of **S**tructure **f**rom **M**otion (SfM). Visual SLAM or vision-based SLAM is a camera-only variant of SLAM which forgoes expensive laser sensors and **i**nertial **m**easurement **u**nits (IMUs). Monocular SLAM uses a single camera while non-monocular SLAM typically uses a pre-calibrated fixed-baseline stereo camera rig. SLAM is prime example of a what is called a "Geometric Method" in Computer Vision. In fact, CMU's Robotics Institute splits the graduate level computer vision curriculum into a Learning-based Methods in Vision course and a separate Geometry-Based Methods in Vision course.

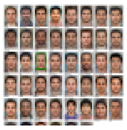
Structure from Motion vs Visual SLAM
Structure from Motion (SfM) and SLAM are solving a very similar problem, but while SfM is traditionally performed in an offline fashion, SLAM has been slowly moving towards the low-power / real-time / single RGB camera mode of operation. Many of the today's top experts in Structure from Motion work for some of the world's biggest tech companies, helping make maps better. Successful mapping products like Google Maps could not have been built without intimate knowledge of multiple-view geometry, SfM, and SLAM. A typical SfM problem is the following: given a large collection of photos of a single outdoor structure (like the Colliseum), construct a 3D model of the structure and determine the camera's poses. The image collection is processed in an offline setting, and large reconstructions can take anywhere between hours and days.

SfM Software: Bundler is one of the most successful SfM open source libraries

Here are some popular SfM-related software libraries:

- Bundler, an open-source Structure from Motion toolkit
- Libceres, a non-linear least squares minimizer (useful for bundle adjustment problems)
- Andrew Zisserman's Multiple-View Geometry MATLAB Functions

Visual SLAM vs Autonomous Driving
While self-driving cars are one of the most important applications of SLAM, according to Andrew Davison, one of the workshop organizers, SLAM for Autonomous Vehicles deserves its own research track. (And as we'll see, none of the workshop presenters talked about self-driving cars). For many years to come it will make sense to continue studying SLAM from a research perspective, independent of any single Holy-Grail



Can a person-specific face recognition algorithm be used to determine a person's race?

It's a valid question: can a person-specific face recognition algorithm be used to determine a person's race? I trained two separa...

Recent Posts

1 **DeepFakes: AI-powered deception machines**

Driven by computer vision and deep learning techniques, a new wave of imaging ... read more

May 16 2018

2 **Nuts and Bolts of Building Deep Learning Applications: Ng @ NIPS2016**

You might go to a cutting-edge machine learning research conference like NIPS ... read more

Dec 16 2016

3 **Making Deep Networks Probabilistic via Test-time Dropout**

In Quantum Mechanics, Heisenberg's Uncertainty Principle states that there is a ... read more

Jun 17 2016

4 **Deep Learning Trends @ ICLR 2016**

Started by the youngest members of the Deep Learning Mafia [1], ... read more

Jun 01 2016

5 **The Future of Real-Time SLAM and Deep Learning vs SLAM**

Last month's International Conference of Computer Vision (ICCV) was full of ... read more

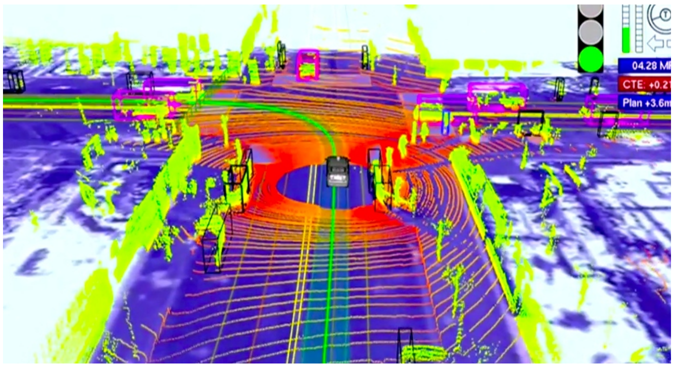
Jan 13 2016

Recent Posts Widget

Links

Tomasz @ MIT Research Homepage

application. While there are just too many system-level details and tricks involved with autonomous vehicles, research-grade SLAM systems require very little more than a webcam, knowledge of algorithms, and elbow grease. As a research topic, Visual SLAM is much friendlier to thousands of early-stage PhD students who'll first need years of in-lab experience with SLAM before even starting to think about expensive robotic platforms such as self-driving cars.



Google's Self-Driving Car's perception system. From IEEE Spectrum's "How Google's Self-Driving Car Works"

Related: March 2015 blog post, Mobileye's quest to put Deep Learning inside every new car.
Related: One way Google's Cars Localize Themselves

Part II: The Future of Real-time SLAM

Now it's time to officially summarize and comment on the presentations from The Future of Real-time SLAM workshop. Andrew Davison started the day with an excellent historical overview of SLAM called 15 years of vision-based SLAM, and his slides have good content for an introductory robotics course.

For those of you who don't know Andy, he is the one and only Professor Andrew Davison of Imperial College London. Most known for his 2003 MonoSLAM system, he was one of the first to show how to build SLAM systems from a single "*monocular"* camera at a time when just everybody thought you needed a stereo "*binocular*" camera rig. More recently, his work has influenced the trajectory of companies such as Dyson and the capabilities of their robotic systems (e.g., the brand new Dyson360).

I remember Professor Davidson from the Visual SLAM tutorial he gave at the BMVC Conference back in 2007. Surprisingly very little has changed in SLAM compared to the rest of the machine-learning heavy work being done at the main vision conferences. In the past 8 years, object recognition has undergone 2-3 mini revolutions, while today's SLAM systems don't look much different than they did 8 years ago. The best way to see the progress of SLAM is to take a look at the most successful and memorable systems. In Davison's workshop introduction talk, he discussed some of these exemplary systems which were produced by the research community over the last 10-15 years:

- **MonoSLAM**
- **PTAM**
- **FAB-MAP**
- **DTAM**
- **KinectFusion**

Davison vs Horn: The next chapter in Robot Vision
Davison also mentioned that he is working on a new Robot Vision book, which should be an exciting treat for researchers in computer vision, robotics, and artificial intelligence. The last Robot Vision book was written by B.K. Horn (1986), and it's about time for an updated take on Robot Vision.

Tomasz @ Google Scholar Citations

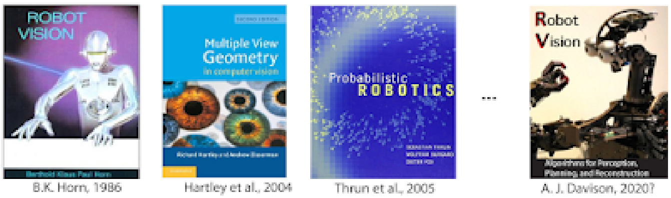
Tomasz @ Github Open-Source Code

Blog Archive

- ► 2018 (1)
- ▼ 2016 (4)
  - ► December (1)
  - ► June (2)
  - ▼ January (1)
    - The Future of Real-Time SLAM and Deep Learning vs ...
- ► 2015 (12)
- ► 2014 (8)
- ► 2013 (13)
- ► 2012 (11)
- ► 2011 (31)
- ► 2010 (24)
- ► 2009 (29)
- ► 2008 (23)
- ► 2007 (11)
- ► 2006 (58)
- ► 2005 (62)

Labels

3d recognition
abhinav gupta
antonio torralba
artificial intelligence
cognitive science
computer vision
cvpr
deep learning
entrepreneurship
future directions
graphical models
iccv
image understanding
MATLAB
MIT

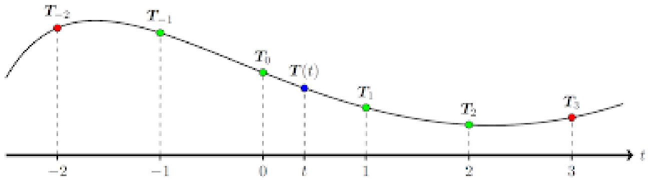

ROBOT VISION
Berthold Klaus Paul Horn
B.K. Horn, 1986
Multiple View Geometry in computer vision
Richard Hartley and Andrew Zisserman
Hartley et al., 2004
Probabilistic ROBOTICS
Sebastian Thrun, Wolfram Burgard, Dieter Fox
Thrun et al., 2005
...
Robot Vision
Algorithms for Perception, Planning, and Reconstruction
A. J. Davison, 2020?

A new robot vision book?

While I’ll gladly read a tome that focuses on the philosophy of robot vision, personally I would like the book to focus on practical algorithms for robot vision, like the excellent Multiple View Geometry book by Hartley and Zissermann or Probabilistic Robotics by Thrun, Burgard, and Fox. A "cookbook" of visual SLAM problems would be a welcome addition to any serious vision researcher's collection.

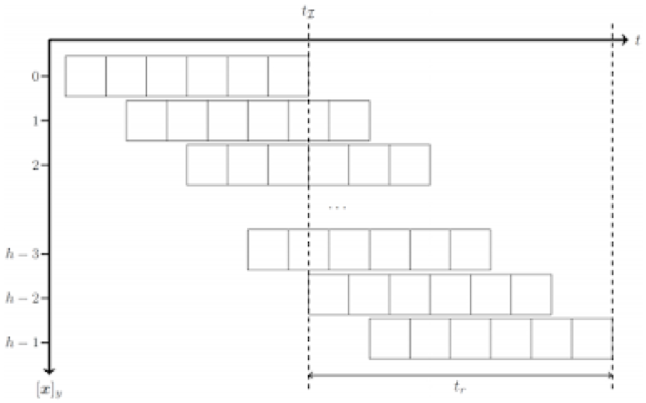
Related: Davison's 15-years of vision-based SLAM slides

Talk 1: Christian Kerl on Continuous Trajectories in SLAM
The first talk, by Christian Kerl, presented a dense tracking method to estimate a continuous-time trajectory. The key observation is that most SLAM systems estimate camera poses at a discrete number of time steps (either they key frames which are spaced several seconds apart, or the individual frames which are spaced approximately 1/25s apart).



Continuous Trajectories vs Discrete Time Points. SLAM/SfM usually uses discrete time points, but why not go continuous?

Much of Kerl’s talk was focused on undoing the damage of rolling shutter cameras, and the system demo’ed by Kerl paid meticulous attention to modeling and removing these adverse rolling shutter effects.



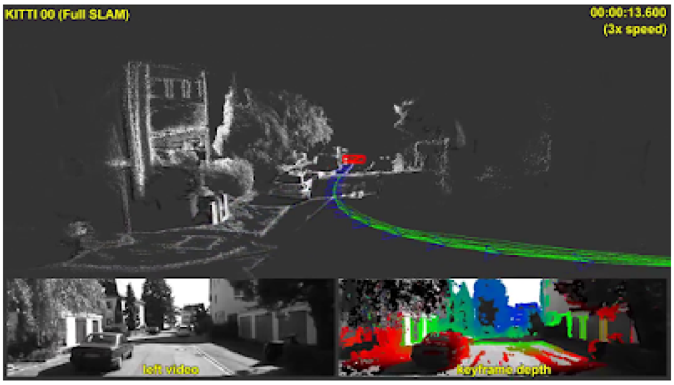
Undoing the damage of rolling shutter in Visual SLAM.

Related: Kerl's Dense continous-time tracking and mapping slides.
Related: Dense Continuous-Time Tracking and Mapping with Rolling Shutter RGB-D Cameras (C. Kerl, J. Stueckler, D. Cremers), In IEEE International Conference on Computer Vision (ICCV), 2015. [pdf]

Talk 2: Semi-Dense Direct SLAM by Jakob Engel
LSD-SLAM came out at ECCV 2014 and is one of my favorite SLAM systems today! Jakob Engel was there to present his system and show the crowd some of the coolest SLAM visualizations in town. LSD-SLAM is an acronym for Large-Scale Direct

nips
object recognition
philosophy
programming
psychology
scene understanding
segmentation
startups
svm
training
visual memex
VMX

Monocular SLAM. LSD-SLAM is an important system for SLAM researchers because it does not use corners or any other local features. **Direct tracking is performed by image-to-image alignment** using a coarse-to-fine algorithm with a robust Huber loss. This is quite different than the feature-based systems out there. Depth estimation uses an inverse depth parametrization (like many other SLAM systems) and uses a large number or relatively small baseline image pairs. Rather than relying on image features, the algorithms is effectively performing “texture tracking”. Global mapping is performed by creating and solving a pose graph "bundle adjustment" optimization problem, and all of this works in real-time. The method is semi-dense because it only estimates depth at pixels solely near image boundaries. LSD-SLAM output is denser than traditional features, but not fully dense like Kinect-style RGBD SLAM.

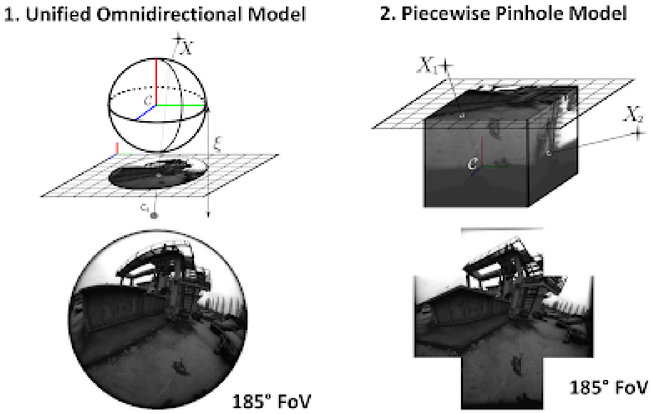


LSD-SLAM in Action: LSD-SLAM generates both a camera trajectory and a semi-dense 3D scene reconstruction. This approach works in real-time, does not use feature points as primitives, and performs direct image-to-image alignment.

Engel gave us an overview of the original LSD-SLAM system as well as a handful of new results, extending their initial system to more creative applications and to more interesting deployments. (See paper citations below)

Related: LSD-SLAM Open-Source Code on github LSD-SLAM project webpage
Related: LSD-SLAM: Large-Scale Direct Monocular SLAM (J. Engel, T. Schöps, D. Cremers), In European Conference on Computer Vision (ECCV), 2014. [pdf] [youtube video]

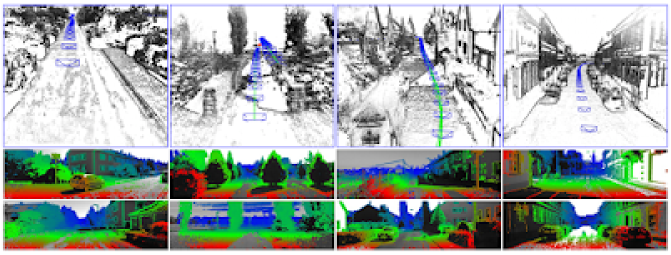
An extension to LSD-SLAM, **Omni LSD-SLAM** was created by the observation that the pinhole model does not allow for a large field of view. This work was presented at IROS 2015 (Caruso is first author) and allows a large field of view (ideally more than 180 degrees). From Engel’s presentation it was pretty clear that you can perform ballerina-like motions (extreme rotations) while walking around your office and holding the camera. This is one of those worst-case scenarios for narrow field of view SLAM, yet works quite well in Omni LSD-SLAM.



Omnidirectional LSD-SLAM Model. See Engel's Semi-Dense Direct SLAM presentation slides.

Related: Large-Scale Direct SLAM for Omnidirectional Cameras (D. Caruso, J. Engel, D. Cremers), In International Conference on Intelligent Robots and Systems (IROS), 2015. [pdf] [youtube video]

Stereo LSD-SLAM is an extension of LSD-SLAM to a binocular camera rig. This helps in getting the absolute scale, initialization is instantaneous, and there are no issues with strong rotation. While monocular SLAM is very exciting from an academic point of view, if your robot is a 30,000$ car or 10,000$ drone prototype, you should have a good reason to not use a two+ camera rig. Stereo LSD-SLAM performs quite competitively on SLAM benchmarks.



Stereo LSD-SLAM. Excellent results on KITTI vehicle-SLAM dataset.

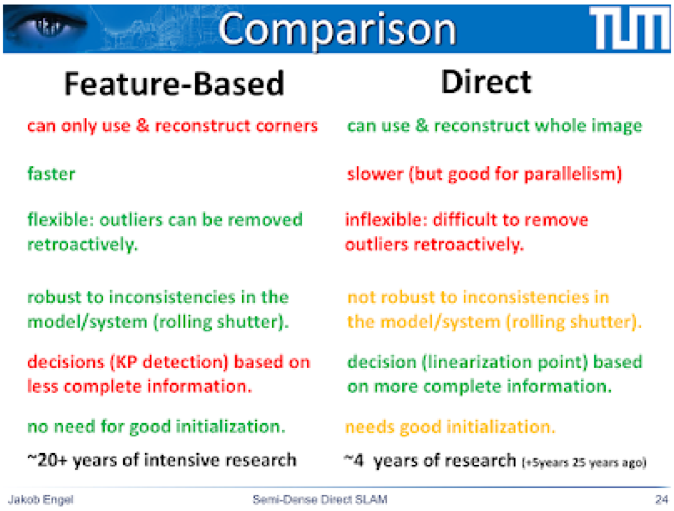
Stereo LSD-SLAM is quite practical, optimizes a pose graph in SE(3), and includes a correction for auto exposure. The goal of auto-exposure correcting is to make the error function invariant to affine lighting changes. The underlying parameters of the color-space affine transform are estimated during matching, but thrown away to estimate the image-to-image error. From Engel's talk, outliers (often caused by over-exposed image pixels) tend to be a problem, and much care needs to be taken to care of their effects.

Related: Large-Scale Direct SLAM with Stereo Cameras (J. Engel, J. Stueckler, D. Cremers), In International Conference on Intelligent Robots and Systems (IROS), 2015. [pdf] [youtube video]

Later in his presentation, Engel gave us a sneak peak on new research about i*ntegrating both stereo and inertial sensors*. For details, you'll have to keep hitting refresh on Arxiv or talk to Usenko/Engel in person. On the applications side, Engel's presentation included updated videos of an Autonomous Quadrotor driven by LSD-SLAM. The flight starts with an up-down motion to get the scale estimate and a free-space octomap is used to estimate the free-space so that the quadrotor can navigate space on its own. Stay tuned for an official publication...

Quadrotor running Stereo LSD-SLAM.
See Engel's quadrotor youtube video from 2012.

The story of LSD-SLAM is also the story of **feature-based vs direct-methods** and Engel gave both sides of the debate a fair treatment. Feature-based methods are engineered to work on top of Harris-like corners, while direct methods use the entire image for alignment. Feature-based methods are faster (as of 2015), but direct methods are good for parallelism. Outliers can be retroactively removed from feature-based systems, while direct methods are less flexible w.r.t. outliners. Rolling shutter is a bigger problem for direct methods and it makes sense to use a global shutter or a rolling shutter model (see Kerl's work). Feature-based methods require making decisions using incomplete information, but direct methods can use much more information. Feature-based methods have no need for good initialization and direct-based methods need some clever tricks for initialization. There is only about 4 years of research on direct methods and 20+ on sparse methods. Engel is optimistic that direct methods will one day rise to the top, and so am I.



Comparison

Feature-Based	Direct
can only use & reconstruct corners	can use & reconstruct whole image
faster	slower (but good for parallelism)
flexible: outliers can be removed retroactively.	inflexible: difficult to remove outliers retroactively.
robust to inconsistencies in the model/system (rolling shutter).	not robust to inconsistencies in the model/system (rolling shutter).
decisions (KP detection) based on less complete information.	decision (linearization point) based on more complete information.
no need for good initialization.	needs good initialization.
~20+ years of intensive research	~4 years of research (+5years 25 years ago)

Jakob Engel Semi-Dense Direct SLAM 24

Feature-based vs direct methods of building SLAM systems. Slide from Engel's talk.

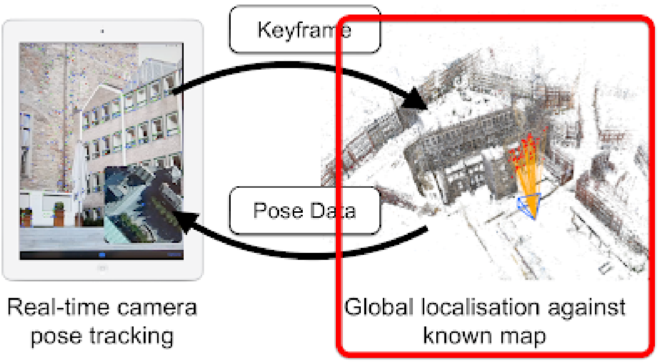
At the end of Engel's presentation, Davison asked about semantic segmentation and Engel wondered whether semantic segmentation can be performed directly on semi-dense "near-image-boundary" data. However, my personal opinion is that there are better ways to apply semantic segmentation to LSD-like SLAM systems. Semi-dense SLAM can focus on geometric information near boundaries, while object recognition can focus on reliable semantics away from the same boundaries, potentially creating a hybrid geometric/semantic interpretation of the image.

Related: Engel's Semi-Dense Direct SLAM presentation slides

Talk 3: Sattler on The challenges of Large-Scale Localization and Mapping
Torsten Sattler gave a talk on large-scale localization and mapping. The motivation for this work is to perform 6-dof localization inside an existing map, especially for mobile localization. One of the key points in the talk was that when you are using traditional feature-based methods, storing your descriptors soon becomes very costly. Techniques such as visual vocabularies (remember product quantization?) can significantly reduce memory overhead, and with clever optimization at some point storing descriptors no longer becomes the memory bottleneck.

Another important take-home message from Sattler's talk is that the number of inliers is not actually a good confidence measure for camera pose estimation. When the feature point are all concentrated in a single part of the image, camera localization can be kilometers away! A better measure of confidence is the "effective inlier count" which looks at the area spanned by the inliers as a fraction of total image area. What you really want is feature matches from all over the image — if the information is spread out across the image you get a much better pose estimate.

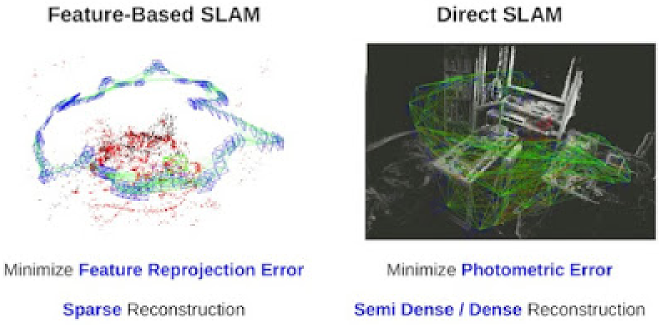
Sattler's take on the future of real-time slam is the following: we should focus on compact map representations, we should get better at understanding camera pose estimate confidences (like down-weighing features from trees), we should work on more challenging scenes (such as worlds with planar structures and nighttime localization against daytime maps).


Keyframe
Pose Data
Real-time camera pose tracking
Global localisation against known map

Mobile Localisation: Sattler's key problem is localizing yourself inside a large city with a single smartphone picture

Related: Scalable 6-DOF Localization on Mobile Devices. Sven Middelberg, Torsten Sattler, Ole Untzelmann, Leif Kobbelt. In ECCV 2014. [pdf]
Related: Torsten Sattler 's The challenges of large-scale localisation and mapping slides

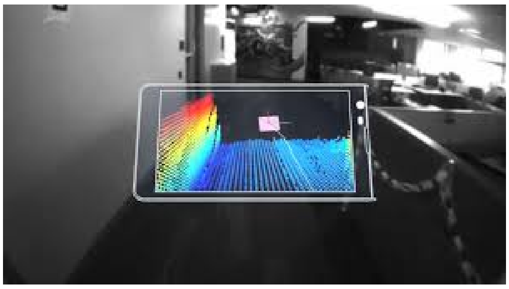
Talk 4: Mur-Artal on Feature-based vs Direct-Methods
Raúl Mur-Artal, the creator of ORB-SLAM, dedicated his entire presentation to the Feature-based vs Direct-method debate in SLAM and he's definitely on the feature-based side. ORB-SLAM is available as an open-source SLAM package and it is hard to beat. During his evaluation of ORB-SLAM vs PTAM it seems that PTAM actually fails quite often (at least on the TUM RGB-D benchmark). LSD-SLAM errors are also much higher on the TUM RGB-D benchmark than expected.


Feature-Based SLAM
Direct SLAM
Minimize Feature Reprojection Error
Sparse Reconstruction
Minimize Photometric Error
Semi Dense / Dense Reconstruction

Feature-Based SLAM vs Direct SLAM. See Mur-Artal's Should we still do sparse feature based SLAM? presentation slides

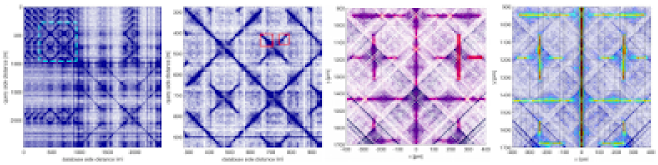
Related: Mur-Artal's Should we still do sparse-feature based SLAM? slides
Related: Monocular ORB-SLAM R. Mur-Artal, J. M. M. Montiel and J. D. Tardos. A versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics. 2015 [pdf]
Related: ORB-SLAM Open-source code on github, Project Website

Talk 5: Project Tango and Visual loop-closure for image-2-image constraints
Simply put, Google's Project Tango is the world' first attempt at commercializing SLAM. Simon Lynen from Google Zurich (formerly ETH Zurich) came to the workshop with a Tango live demo (on a tablet) and a presentation on what's new in the world of Tango. In case you don't already know, Google wants to put SLAM capabilities into the next generation of Android Devices.



Google's Project Tango needs no introduction.

The Project Tango presentation discussed a new way of doing loop closure by finding certain patters in the image-to-image matching matrix. This comes from the “Placeless Place Recognition” work. They also do online bundle adjustment w/ vision-based loop closure.



Loop Closure inside a Project Tango? Lynen et al's Placeless Place Recognition. The image-to-image matrix reveals a new way to look for loop-closure. See the algorithm in action in this youtube video.

The Project Tango folks are also working on combing multiple crowd-sourced maps at Google, where the goals to combine multiple mini-maps created by different people using Tango-equipped devices.

Simon showed a video of mountain bike trail tracking which is actually quite difficult in practice. The idea is to go down a mountain bike trail using a Tango device and create a map, then the follow-up goal is to have a separate person go down the trail. This currently “semi-works” when there are a few hours between the map building and the tracking step, but won’t work across weeks/months/etc.

During the Tango-related discussion, Richard Newcombe pointed out that the “features” used by Project Tango are quite primitive w.r.t. getting a deeper understanding of the environment, and it appears that Project Tango-like methods won't work on outdoor scenes where the world is plagued by non-rigidity, massive illumination changes, etc. So are we to expect different systems being designed for outdoor systems or will Project Tango be an indoor mapping device?

Related: Placeless Place Recognition. Lynen, S. ; Bosse, M. ; Furgale, P. ; Siegwart, R. In 3DV 2014.
Related: Google I/O talk from May 29, 2015 about Tango

Talk 6: ElasticFusion is DenseSLAM without a pose-graph
ElasticFusion is a dense SLAM technique which requires a RGBD sensor like the Kinect. 2-3 minutes to obtain a high-quality 3D scan of a single room is pretty cool. A pose-graph is used behind the scenes of many (if not most) SLAM systems, and this technique has a different (map-centric) approach. The approach focuses on building a map, but the trick is that the map is deformable, hence the name ElasticFusion. The “Fusion” part of the algorithm is in homage to KinectFusion which was one of the first high quality kinect-based reconstruction pipelines. Also surfels are used as the underlying primitives.

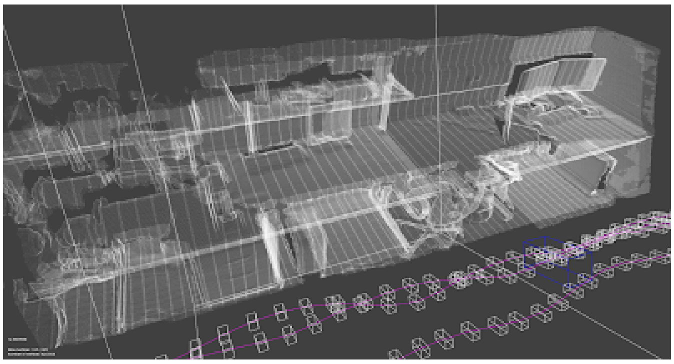
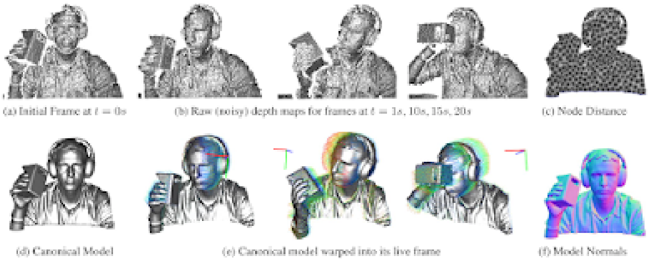


Image from Kintinuous, an early version of Whelan's Elastic Fusion.

Recovering light sources: we were given a sneak peak at new unpublished work from Imperial College London / dyson Robotics Lab. The idea is that detecting the light source direction and detecting specularities, you can improve 3D reconstruction results. Cool videos of recovering light source locations which work for up to 4 separate lights.

Related: Map-centric SLAM with ElasticFusion presentation slides
Related: ElasticFusion: Dense SLAM Without A Pose Graph. Whelan, Thomas and Leutenegger, Stefan and Salas-Moreno, Renato F and Glocker, Ben and Davison, Andrew J. In RSS 2015.

Talk 7: Richard Newcombe's DynamicFusion
Richard Newcombe's (whose recently formed company was acquired by Oculus), was the last presenter. It's really cool to see the person behind DTAM, KinectFusion, and DynamicFusion now working in the VR space.

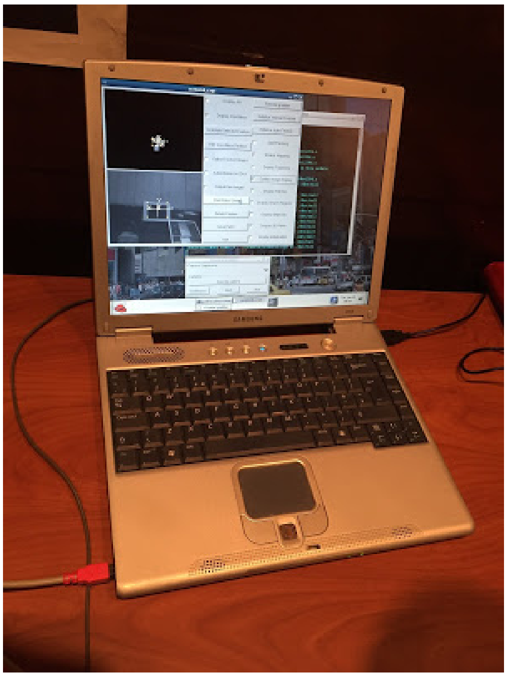

(a) Initial Frame at $t = 0s$
(b) Raw (noisy) depth maps for frames at $t = 1s, 10s, 15s, 20s$
(c) Node Distance
(d) Canonical Model
(e) Canonical model warped into its live frame
(f) Model Normals

Newcombe's Dynamic Fusion algorithm. The technique won the prestigious CVPR 2015 best paper award, and to see it in action just take a look at the authors' DynamicFusion Youtube video.

Related: DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time, Richard A. Newcombe, Dieter Fox, Steven M. Seitz. In CVPR 2015. [pdf] [Best-Paper winner]
Related: SLAM++: Simultaneous Localisation and Mapping at the Level of Objects Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H. J. Kelly and Andrew J. Davison (CVPR 2013)
Related: KinectFusion: Real-Time Dense Surface Mapping and Tracking Richard A. Newcombe Shahram Izadi,Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Andrew Fitzgibbon (ISMAR 2011, Best paper award!)

Workshop Demos
During the demo sessions (held in the middle of the workshop), many of the presenter showed off their SLAM systems in action. Many of these systems are available as open-source (free for non-commercial use?) packages, so if you're interested in real-time SLAM, downloading the code is worth a shot. However, **the one demo which stood out was Andrew Davison's showcase of his MonoSLAM system from 2004**. Andy had to revive his 15-year old laptop (which was running Redhat Linux) to show off his original

system, running on the original hardware. If the computer vision community is going to oneway decide on a “retro-vision” demo session, I’m just going to go ahead and nominate Andy for the best-paper prize, right now.



Andry's Retro-Vision SLAM Setup (Pictured on December 18th, 2015)

It was interesting to watch the SLAM system experts wave their USB cameras around, showing their systems build 3D maps of the desk-sized area around their laptops. If you carefully look at the way these experts move the camera around (i.e., smooth circular motions), you can almost tell how long a person has been working with SLAM. When the non-experts hold the camera, probability of tracking failure is significantly higher.

I had the pleasure of speaking with Andy during the demo session, and I was curious which line of work (in the past 15 years) surprised him the most. His reply was that PTAM, which showed how to perform real-time bundle adjustment, surprised him the most. The PTAM system was essentially a MonoSLAM++ system, but the significantly improved tracking results were due to taking a heavyweight algorithm (bundle adjustment) and making it real-time — something which Andy did not believe was possible in the early 2000s.

Part III: Deep Learning vs SLAM

The SLAM panel discussion was a lot of fun. Before we jump to the important Deep Learning vs SLAM discussion, I should mention that each of the workshop presenters agreed that **semantics are necessary to build bigger and better SLAM systems**. There were lots of interesting mini-conversations about future directions. During the debates, Marc Pollefeys (a well-known researcher in SfM and Multiple-View Geometry) reminded everybody that **Robotics is the killer application of SLAM** and suggested we keep an eye on the prize. This is quite surprising since SLAM was traditionally applied to Robotics problems, but the lack of Robotics success in the last few decades (Google Robotics?) has shifted the focus of SLAM away from Robots and towards large-scale map building (ala Google Maps) and Augmented Reality. Nobody at this workshop talked about Robots.

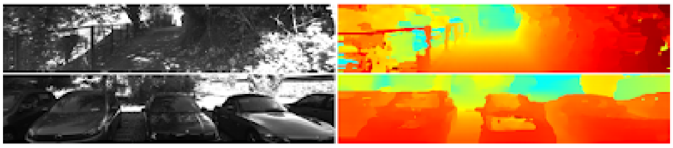
Integrating semantic information into SLAM

There was a lot of interest in incorporating semantics into today’s top-performing SLAM systems. When it comes to semantics, the **SLAM community is unfortunately stuck in the world of bags-of-visual-words**, and doesn't have new ideas on how to integrate semantic information into their systems. On the other end, we’re now seeing real-time semantic segmentation demos (based on ConvNets) popping up at CVPR/ICCV/ECCV, and in my opinion SLAM needs Deep Learning as much as the other way around.

Integrating semantics into SLAM is often talk about, but it is easier said than done. Figure 6.9 (page 142) from Moreno's PhD thesis: Dense Semantic SLAM

"Will end-to-end learning dominate SLAM?"
Towards the end of the SLAM workshop panel, Dr. Zeeshan Zia asked a question which *startled* the entire room and led to a memorable, energy-filled discussion. You should have seen the look on the panel's faces. It was a bunch of geometers being thrown a fireball of deep learning. Their facial expressions suggest both bewilderment, anger, and disgust. "*How dare you question us?*" they were thinking. And it is only during these fleeting moments that we can truly appreciate the conference experience. Zia's question was essentially: **Will end-to-end learning soon replace the mostly manual labor involved in building today's SLAM systems?**.

Zia's question is very important because end-to-end trainable systems have been slowly creeping up on many advanced computer science problems, and there's no reason to believe SLAM will be an exception. A handful of the presenters pointed out that current SLAM systems rely on too much geometry for a pure deep-learning based SLAM system to make sense -- we should use learning to make the point descriptors better, but leave the geometry alone. *Just because you can use deep learning to make a calculator, it doesn't mean you should.*



Learning Stereo Similarity Functions via ConvNets, by Yan LeCun and collaborators.

While many of the panel speakers responded with a somewhat affirmative "no", it was Newcombe which surprisingly championed what the marriage of Deep Learning and SLAM might look like.

Newcombe's Proposal: Use SLAM to fuel Deep Learning
Although Newcombe didn't provide much evidence or ideas on how Deep Learning might help SLAM, he provided **a clear path on how SLAM might help Deep Learning.** Think of all those maps that we've built using large-scale SLAM and all those correspondences that these systems provide — isn't that a clear path for building terascale image-image "association" datasets which should be able to help deep learning? The basic idea is that today's SLAM systems are large-scale "correspondence engines" which can be used to generate large-scale datasets, precisely what needs to be fed into a deep ConvNet.

Concluding Remarks
There is quite a large disconnect between the kind of work done at the mainstream ICCV conference (heavy on machine learning) and the kind of work presented at the real-time

SLAM workshop (heavy on geometric methods like bundle adjustment). The mainstream Computer Vision community has witnessed several mini-revolutions within the past decade (e.g., Dalal-Triggs, DPM, ImageNet, ConvNets, R-CNN) while the SLAM systems of today don't look very different than they did 8 years ago. The Kinect sensor has probably been the single largest game changer in SLAM, but the fundamental algorithms remain intact.

Integrating semantic information: The next frontier in Visual SLAM.
Brain image from Arwen Wallington's blog post.

Today's SLAM systems help machines geometrically understand the immediate world (i.e., build associations in a local coordinate system) while today's Deep Learning systems help machines reason categorically (i.e., build associations across distinct object instances). In conclusion, I share Newcombe and Davison excitement in Visual SLAM, as vision-based algorithms are going to turn Augmented and Virtual Reality into billion dollar markets. However, we should not forget to keep our eyes on the "trillion-dollar" market, the one that's going to redefine what it means to "work" -- namely *Robotics*. The day of Robot SLAM will come soon.

Posted by Unknown at Wednesday, January 13, 2016

Labels: andrew davison, bundle adjustment, DTAM, DynamicFusion, iccv 2015, jakob engel, KinectFusion, LSD-SLAM, marc pollefeys, pose, PTAM, real-time, richard newcombe, robotics, segmentation, sfm, SLAM, workshop, zisserman

27 comments:

Unknown 11:52 PM

What an incredible posting!!

Reply

Anonymous 3:24 PM

Thanks for sharing. This blog post is awesome!

Reply

Anonymous 8:51 AM

A related workshop: "The Problem of Mobile Sensors: Setting future goals and indicators of progress for SLAM" http://ylatif.github.io/movingsensors/

It also has a Google+ community:
https://plus.google.com/u/0/communities/102832228492942322585

Reply

Replies

T Tùng Nguyễn 11:57 AM

Thanks a lot!

Reply

Anonymous 11:20 PM

Great post, i always read your blog, keep it updating !! Kind regards from chile

Reply

Howard 4:49 AM

Waht a great survey! should be published in robotics community.

Reply

Ankur 1:24 PM

Thanks for writing this.

Reply

Ro Tr 2:39 PM

Great overview! Looks like I will really have to compare ORB vs LSD..

Reply

T Tùng Nguyễn 11:53 AM

Is 2020 the planned date of the new Robot Vision book? Can't wait!

Reply

Unknown 12:57 PM

Thanks a lot for this post, very interesting

Reply

Unknown 8:29 AM

Simply brilliant.

Reply

Unknown 8:30 AM

Simply brilliant blog post.

Reply

Anonymous 9:37 PM

Perfect overview on today's SLAM and its community. I will introduce this post to my colleagues. Many thanks!

Reply

Anonymous 4:45 AM

nice

Reply

Djela 9:09 AM

Coming from a filmmaking and UX design background, I understand more and more why I often end up on this blog. Many thanks! Hope a get a chance to try your VMX software soon. All the best.

Reply

zebrajack 5:41 AM

thanks for sharing.

Reply

Anonymous 7:58 PM

Thanks for writing .

Reply

Anonymous 3:06 PM

Awesome, thank you very much!!!

Reply

Jim 3:34 PM

That was a refreshing read. Your blog posts are amazing, keep it up!

Reply

Mary McClelland 9:44 AM

Nice

Reply

阿飞 8:13 AM

It is great post. How will the SLAM help DL is very interesting topic, is there any further materials?

Reply

Danish 4:32 PM

Love this!

Reply

Manohar Kuse 2:32 AM

Fantastic insights and much deep insights. Thanks for sharing...!

Reply

piedad 7:30 AM

great

Reply

Stasya 7:51 AM

Hey,nice

Reply

Unknown 3:19 AM

Thanks for it!

Reply

Ramkumar 4:12 AM

Thank you very much.Great work !

Reply

Enter your comment...

Comment as: kipochenuestc(▼ Sign out

Publish Preview □ Notify me

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)

Follow by Email

Email address... Submit

Subscribe To

Posts

Comments

Awesome Inc. theme. Powered by Blogger.