DEVIEW
2019

When Deep Learning meets Visual Localization

Martin Humenberger
NAVER LABS Europe, 3D Vision Group
http://europe.naverlabs.com

NAVER LABS
EUROPE

Outline

1. 3D Vision @ NAVER LABS Europe
2. Visual Localization: Concept, Methods, Datasets
3. Local Feature Extraction (R2D2)
4. VSLAM in Dynamic Environments (Slamantic)

DEVIEW
2019


GRENOBLE, FRANCE
United Kingdom
Ireland
Poland
Germany
France
Italy
Spain
Portugal

DEVIEW 2019

3D VISION at NAVER LABS Europe

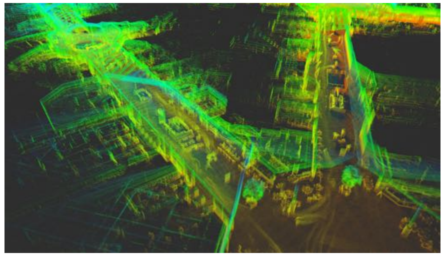
3D Vision – Research Interests

DEVIEW 2019

We want to overcome current limitations of traditional, mainly geometry-based, methods of 3D vision using data driven machine learning techniques.

Main research topics:

a) *Fundamental methods of 3D vision*
   - Correspondence analysis
   - Depth estimation
b) *Camera pose estimation*
   - Visual localization
   - VSLAM / VO
c) *3D scene understanding*
   - Semantic mapping
   - 3D reconstruction
d) Synthetic datasets and domain adaptation
   - Transfer between synthetic and real world



DEVIEW
2019

Visual Localization

Visual Localization - Concept

Château de Sceaux

GPS accuracy sometimes not enough.

E.g. for precise robot navigation or augmented reality.

Goal: Use an image to estimate the precise position of the camera within a given area (map).

Visual Localization - Concept

DEVIEW 2019

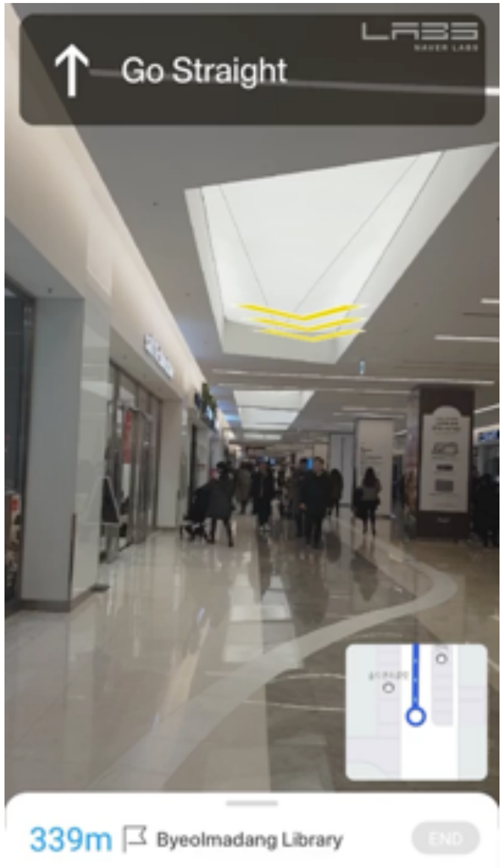
This works indoor as well!

There, it is in particular useful since GPS is not available.

Visual Localization

Application Examples

DEVIEW 2019



Google

NAVER LABS

iRobot

DEVIEW
2019

Methods of Visual Localization

Challenges of Visual Localization


reference image (map)


viewpoint and scale


illumination


occlusion

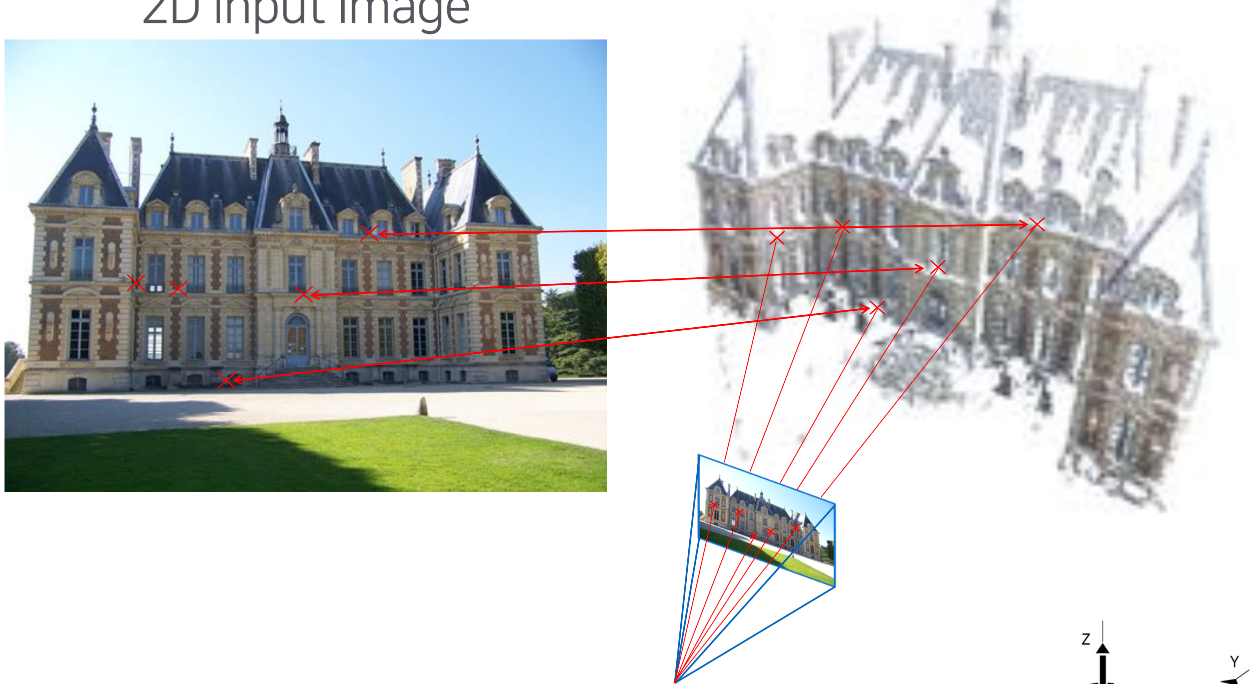

viewpoint, occlusion, weather

Structure Based Visual Localization

Take a picture

2D Input Image

3D Map



Input image → Feature detection & description → Descriptor matching to get 2D-3D correspondences → Camera pose estimation → Location

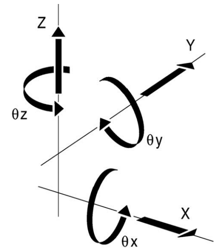
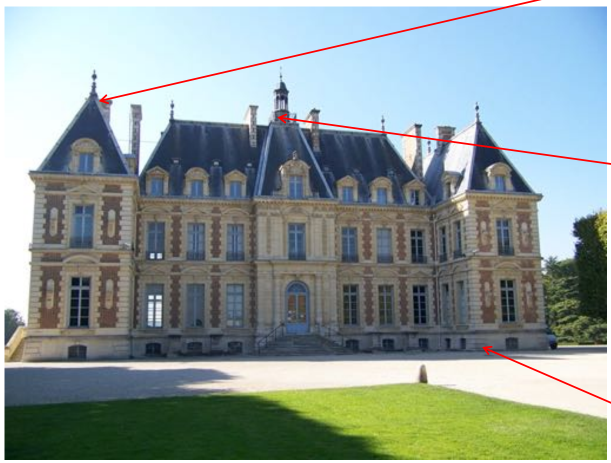
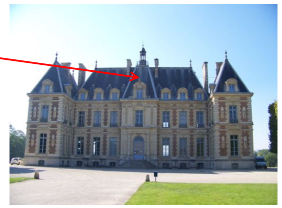


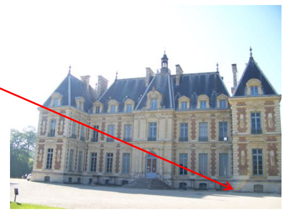
Image Retrieval Based Visual Localization

DEVIEW 2019

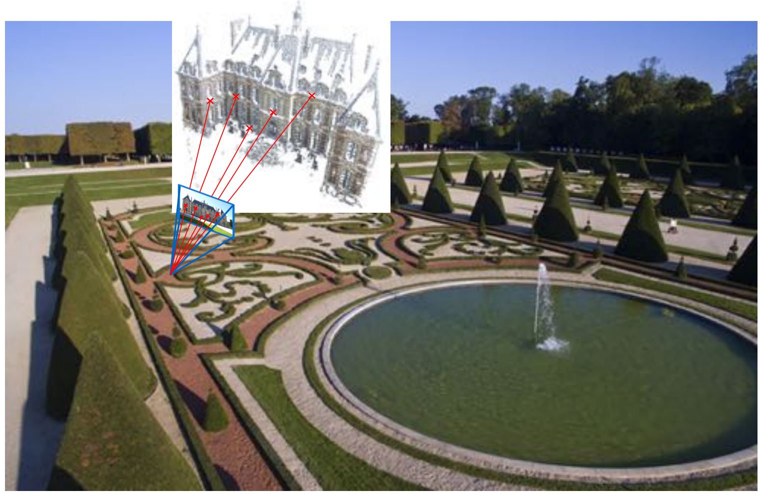
2D Input Image



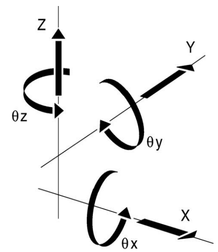




Large 3D Map



Input image → Image retrieval → Descriptor matching to get 2D-3D correspondences → Camera pose estimation → Location



Camera Pose Regression Based Visual Localization

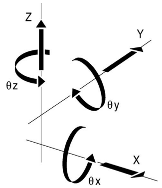


image camera pose

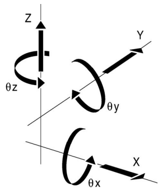


image camera pose

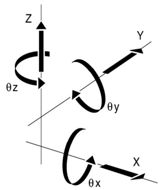
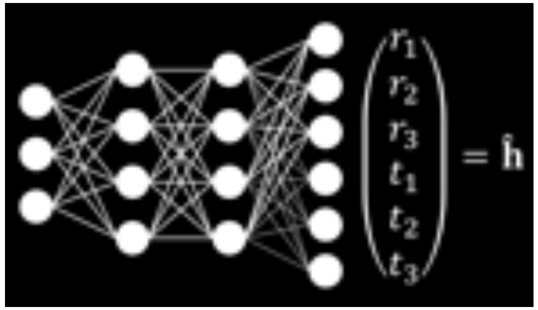


image camera pose

No 3D Map

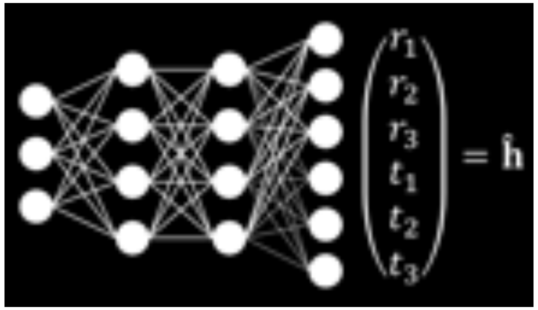
but…



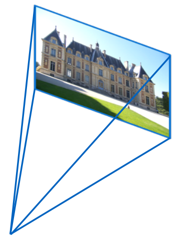
CNN

Camera Pose Regression Based Visual Localization

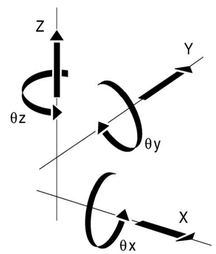
2D Input Image



CNN



Input image → CNN to directly estimate the camera pose → Location



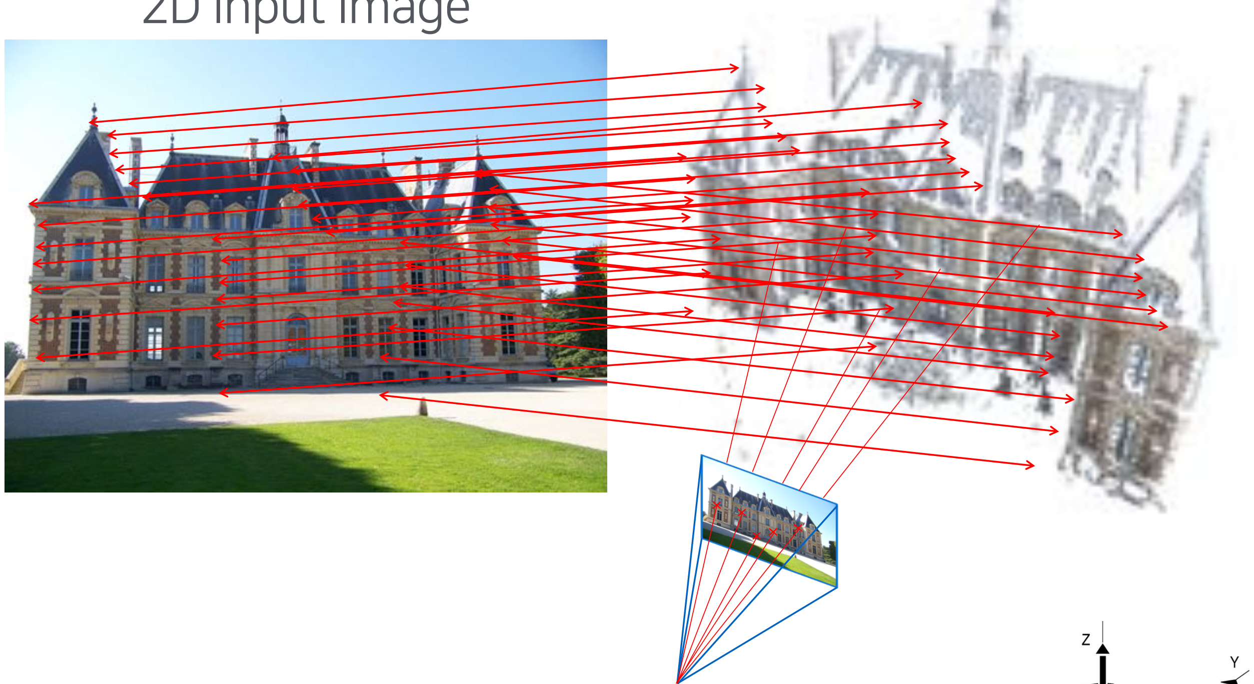
Scene Coordinate Regression Based Visual Localization

DEVIEW 2019

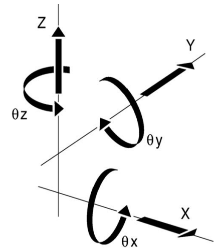
Take a picture

2D Input Image

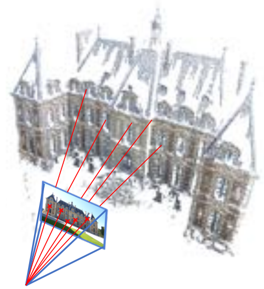
3D Map



Input image → ~~Feature detection & description~~ → CNN to regress dense 2D-3D correspondences → Camera pose estimation → Location



Overview of Methods



DEVIEW 2019

Structure-based methods	Active Search [1] OpenMVG [2]	+ Perform very well on most datasets -> high accuracy - Not suitable for very large environments (memory and processing time)
Image retrieval-based methods	HF-Net [3]	+ Improve speed and robustness for large scale settings - Quality heavily relies on image retrieval
Camera pose regression methods	PoseNet [4]	+ Interesting approach because no 3D maps are needed and it is data driven (can be trained for certain challenges) - Low accuracy
Scene coordinate regression methods	DSAC++ [5]	+ Very accurate in small scale settings - Does not yet work in large scale environments

[1] T. Sattler et al., Improving Image-Based Localization by Active Correspondence Search, ECCV 2012

[2] P. Moulon, OpenMVG: http://github.com/openMVG/openMVG

[3] Sarlin et al., From Coarse to Fine: Robust Hierarchical Localization at Large Scale, CVPR 2019

[4] A. Kendall et al., PoseNet: http://mi.eng.cam.ac.uk/projects/relocalisation/, ICCV 2015

[5] E. Brachmann et al., Learning Less is More – 6D Camera Localization via 3D Surface Regression, CVPR 2018

DEVIEW
2019

Mapping with M1X

NAVER LABS Mapping Robot M1X

M1 (2016)

M1X (2019)

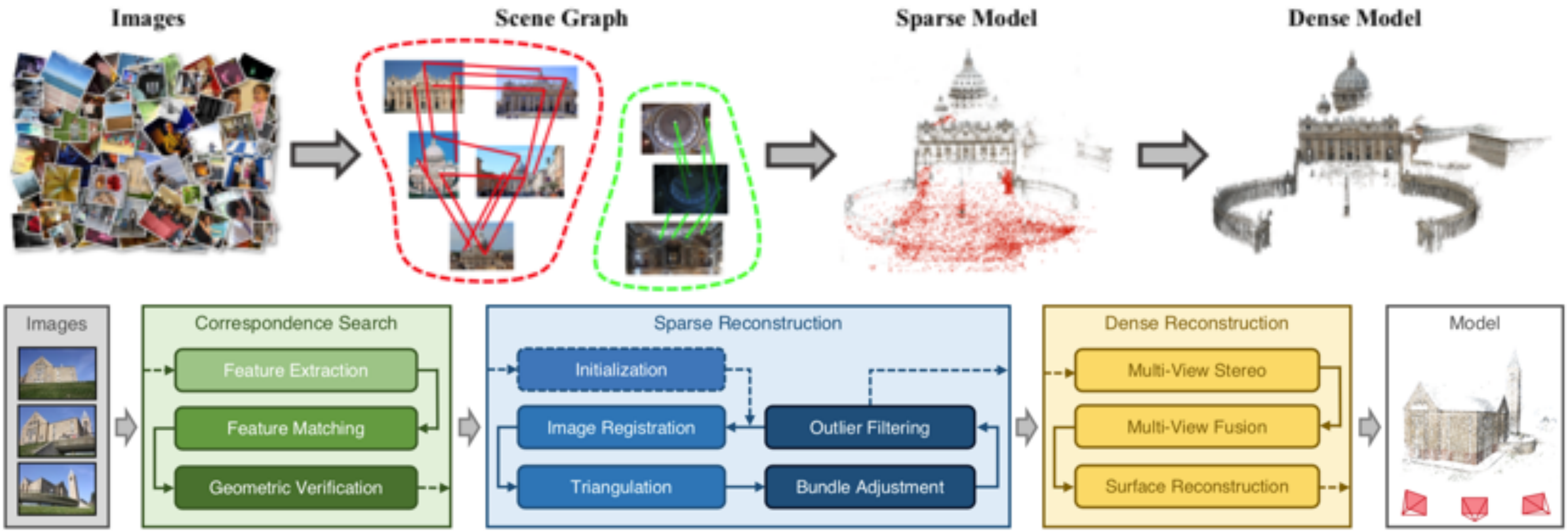

M1

DEVIEW 2019

Mapping with Structure from Motion

Structure from Motion

DEVIEW 2019



J. Schönberger, Robust Methods for Accurate and Efficient 3D Modeling from Unstructured Imagery, PhD, ETHZ

Structure from Motion

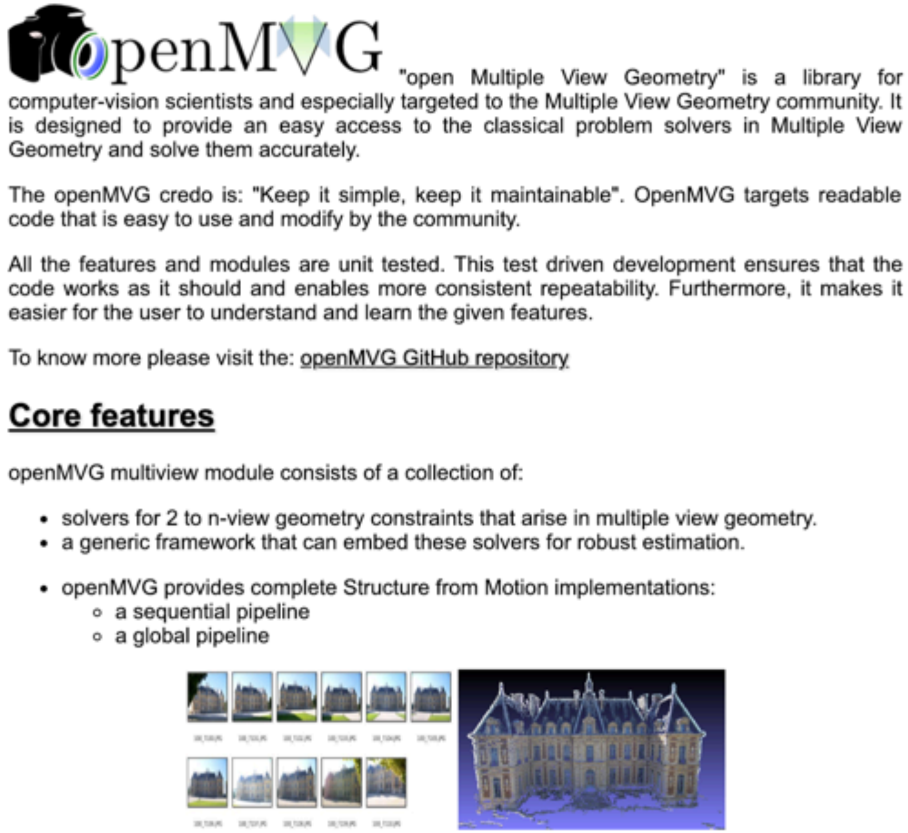
COLMAP – SfM and MVS

About Documentation Download Tutorial

About

COLMAP is a general-purpose Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline with a graphical and command-line interface. It offers a wide range of features for reconstruction of ordered and unordered image collections. The software is licensed under the GNU General Public License. If you use this project for your research, please cite the papers: Paper / Bibtex / Video and Paper / Bibtex / Video.

https://demuc.de/colmap/



openMVG "open Multiple View Geometry" is a library for computer-vision scientists and especially targeted to the Multiple View Geometry community. It is designed to provide an easy access to the classical problem solvers in Multiple View Geometry and solve them accurately.

The openMVG credo is: "Keep it simple, keep it maintainable". OpenMVG targets readable code that is easy to use and modify by the community.

All the features and modules are unit tested. This test driven development ensures that the code works as it should and enables more consistent repeatability. Furthermore, it makes it easier for the user to understand and learn the given features.

To know more please visit the: openMVG GitHub repository

Core features

openMVG multiview module consists of a collection of:

- solvers for 2 to n-view geometry constraints that arise in multiple view geometry.
- a generic framework that can embed these solvers for robust estimation.

- openMVG provides complete Structure from Motion implementations:
  - a sequential pipeline
  - a global pipeline

http://imagine.enpc.fr/~moulonp/openMVG/

DEVIEW 2019

Datasets

Datasets – Cambridge Landmarks – Outdoor Localization

- 8,000 images from 6 scenes up to 100 x 500m RGB, SfM

Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

Slide credit Alex Kendall, https://pdfs.semanticscholar.org/4fc6/7b4dc62e9c8eee4259c3878b71c64958c373.pdf

Datasets – Seven Scenes – Indoor Localization

DEVIEW 2019

- 17,000 images across 7 small indoor scenes. RGB-D, pose, dense reconstruction

Jamie Shotton et al. Scene coordinate regression forests for camera relocalization in RGB-D images. CVPR 2013

Slide credit Alex Kendall, https://pdfs.semanticscholar.org/4fc6/7b4dc62e9c8eee4259c3878b71c64958c373.pdf

Aachen Day-Night

Old inner city of Aachen, Germany
- 4328 reference images
- 922 query images (824 daytime, 98 nighttime)
- All images are captured with hand-held cameras

3D reconstruction (sfm)

training image examples

test image examples (day - night)

Baidu IBL Dataset

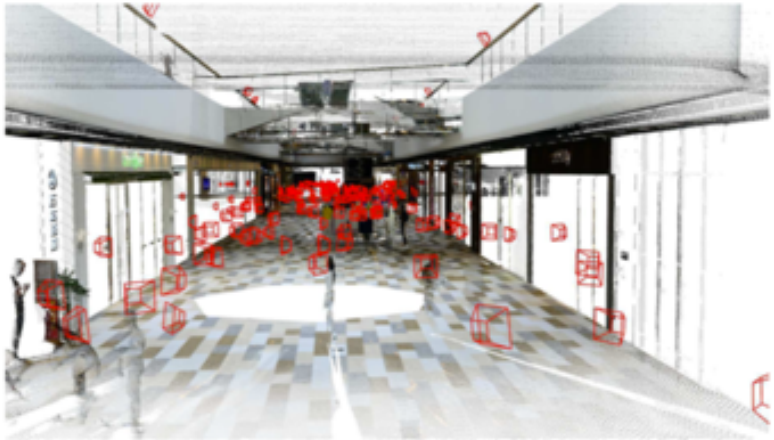
DEVIEW 2019

(a) Captured point cloud in bird-eye view.

- Captured in a shopping mall using high res cameras and a lidar scanner
- RGB (training and testing), point clouds, poses

Sun et al., A Dataset for Benchmarking Image-Based Localization, CVPR17

(b) Close-up of the camera poses for database images.



(c) Groundtruth camera poses for the query images.

Virtual Gallery – Synthetic Dataset

Virtual Gallery – Synthetic Dataset

DEVIEW 2019

Tailored to test specific challenges of visual localization, such as:
- Different lighting conditions
- Occlusions
- Various camera parameters

Training:
- Imitate a robot scanning the museum
- 6 cameras (360º), 1 virtual lidar
- 5 trajectories

Testing:
- Imitate pictures taken by people
- Cameras : Random intrinsics, random orientation, random position
- Different lighting conditions and occlusions

Download: https://europe.naverlabs.com/research/3d-vision/virtual-gallery-dataset/

DEVIEW 2019

Visual Localization using Objects of Interest

Visual Localization by Learning Objects-of-Interest Dense Match Regression

DEVIEW 2019

Published at CVPR19

Philippe Weinzaepfel Gabriela Csurka Yohann Cabon Martin Humenberger
NAVER LABS Europe

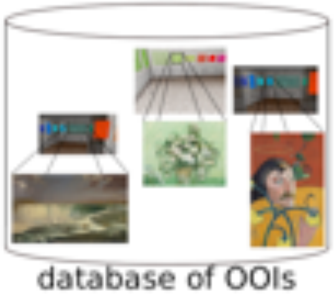
Objects of Interest (OOI) are distinctive areas within the environment which can be detected under various conditions.

Visual Localization by Learning Objects-of-Interest Dense Match Regression

Published at CVPR19

Philippe Weinzaepfel Gabriela Csurka Yohann Cabon Martin Humenberger
NAVER LABS Europe

Map =

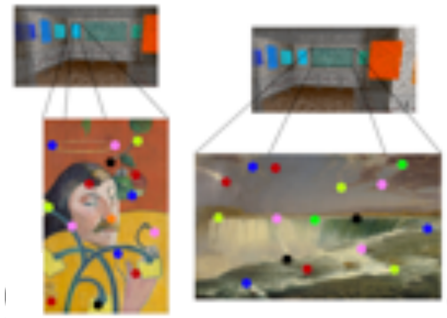


- list of all OOIs
- 3D locations of OOIs

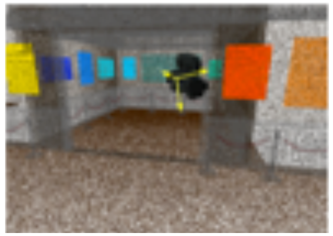
<u>Main advantage:</u>

Data driven approach which can overcome common VL challenges.

1) Start with input image



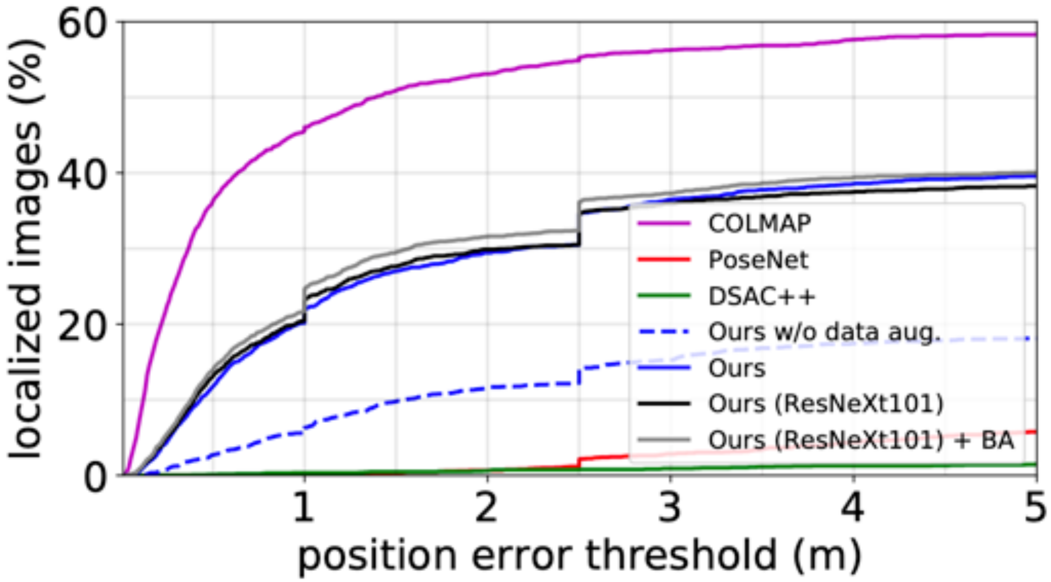
2) Feed into OOI network



3) Use correspondences to compute the camera location

Localization Results – Baidu Dataset

DEVIEW 2019



- Structure-based methods perform best.
- Learning-based methods (PoseNet, DSAC++) do not work on this large dataset.
- Our approach is the first learning-based method which can be applied here.

Paper: https://europe.naverlabs.com/research/publications/visual-localization-by-learning-objects-of-interest-dense-match-regression/

DEVIEW 2019

Local Feature Extraction

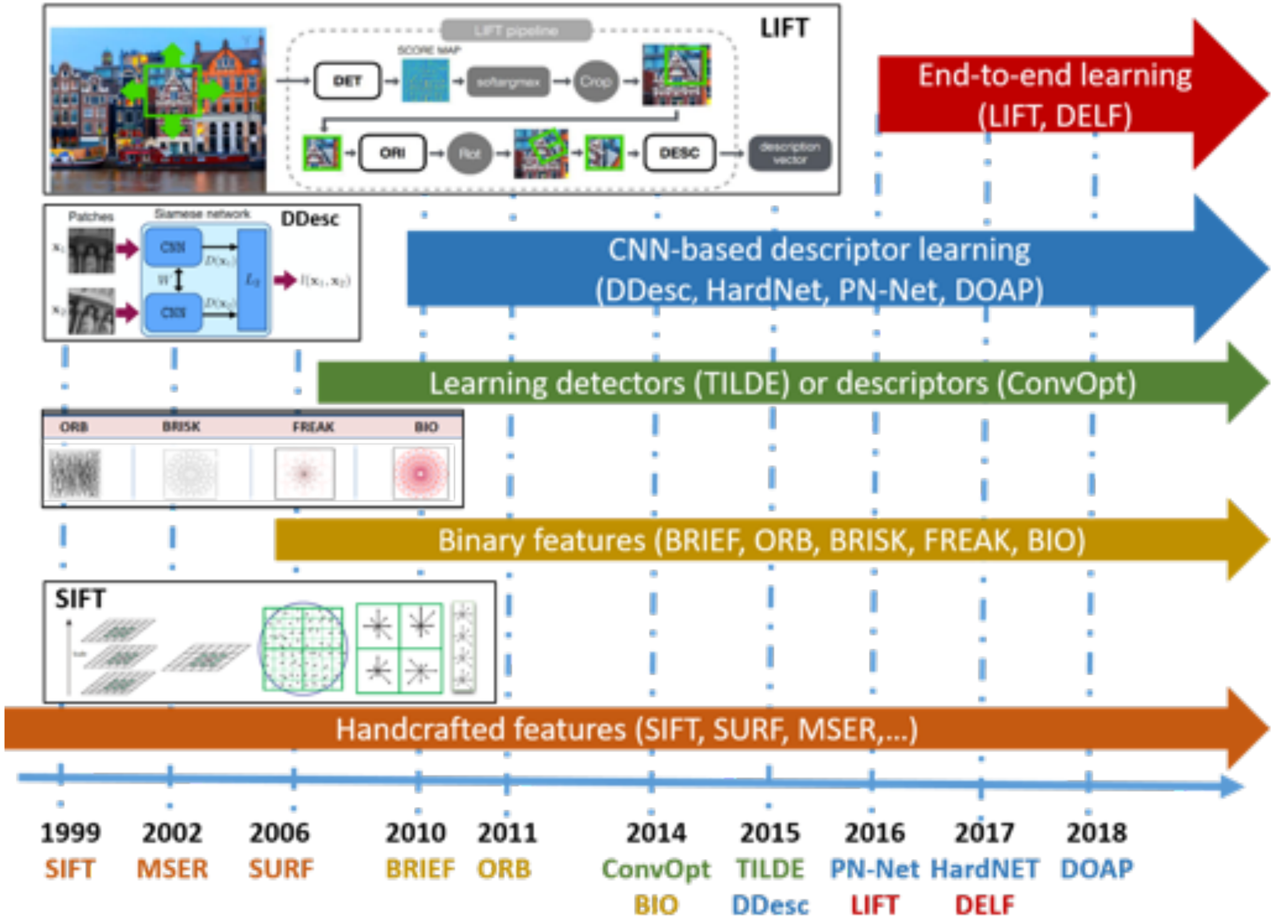
R2D2 – Repeatable and Reliable Detector and Descriptor

Motivation

- Structure-based methods perform well and the critical part is feature extraction and matching.
- A robust feature detector enables robust visual localization
- … and improves many other applications such as object detection, VSLAM and SfM.

Overview

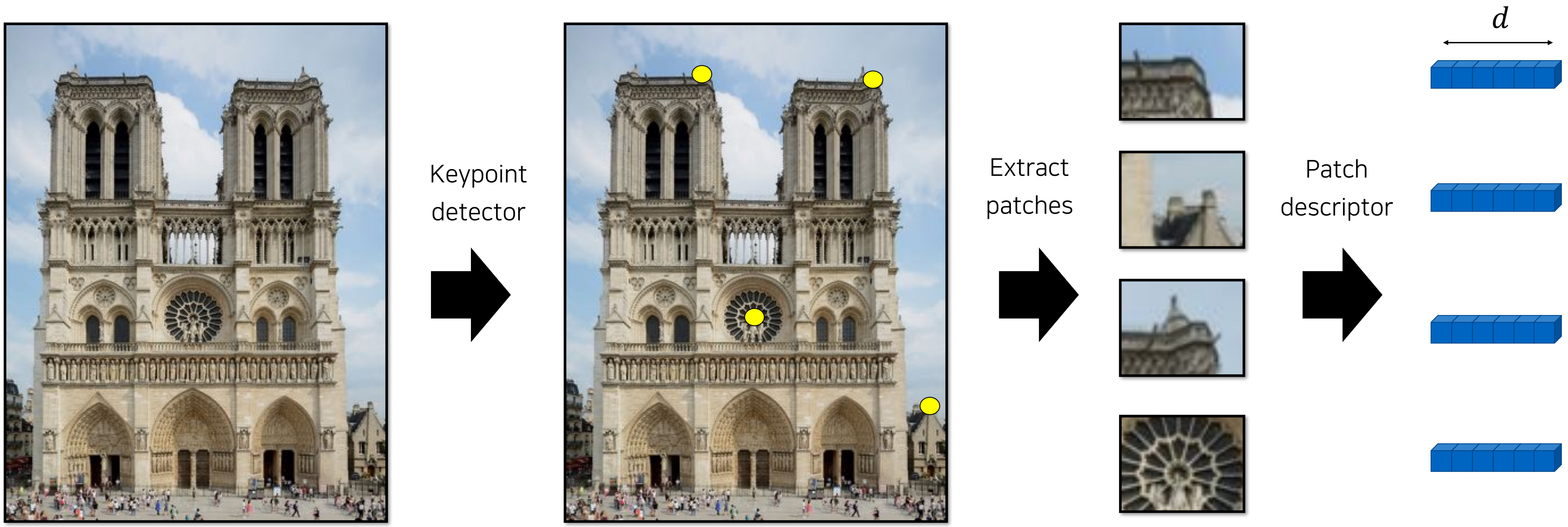
DEVIEW 2019



Csurka et al., From handcrafted to deep local features, arXiv 2018

Introduction - Classical methods: *Detect-**then**-describe*

DEVIEW 2019



1) Start with input image

2) Detect keypoints

3) Describe keypoints

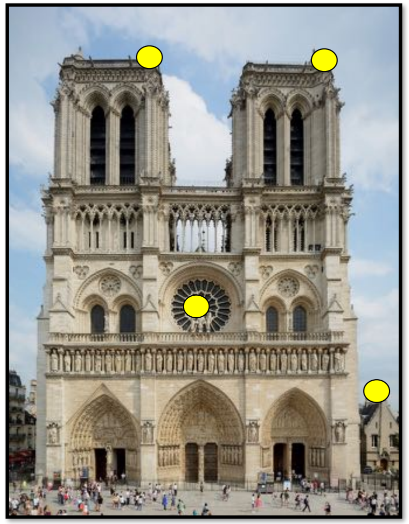
Introduction

- Classical methods: *Detect-**then**-describe*
- Our approach: *Detect-**and**-describe*

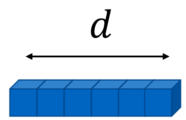
1) Start with input image

2) Feed into R2D2 network

Keypoints
(nms)



descriptor
for each keypoint



3) Detect keypoints &
describe them at once

Approach

- **Repeatability**: image locations that are invariant to usual image transformations (e.g. corners)
- **Reliability**: image locations that are good (discriminative and robust) for matching purpose

Reliable?

Repeatable?

	No	Yes
No	①	②
Yes	③	④

➔ *All cases are possible:*
reliability and repeatability
are independent

Our approach

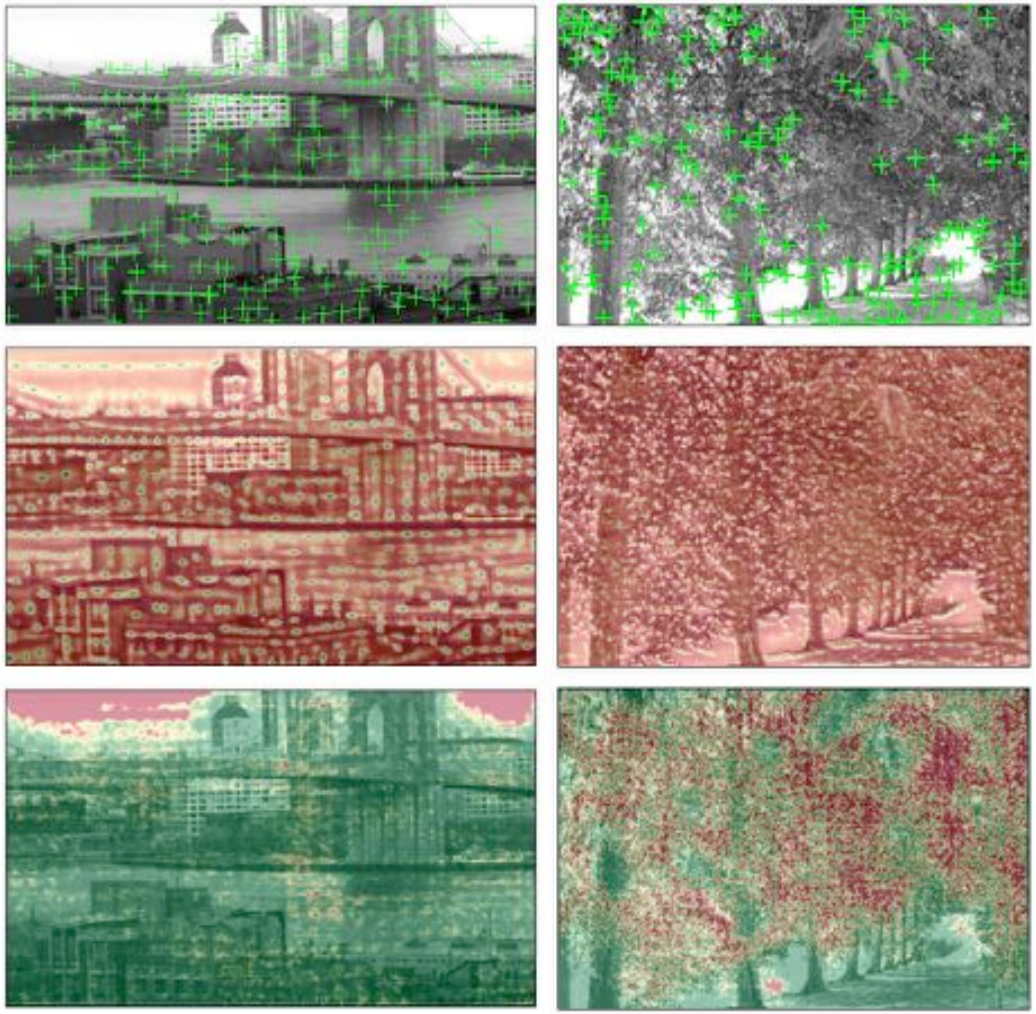
- Detect-and-describe (dense) to predict repeatability and reliability separately
- Novel loss to estimate the reliability (or "matchability")
- Novel self-supervised loss to learn repeatability without introducing any biases

Results

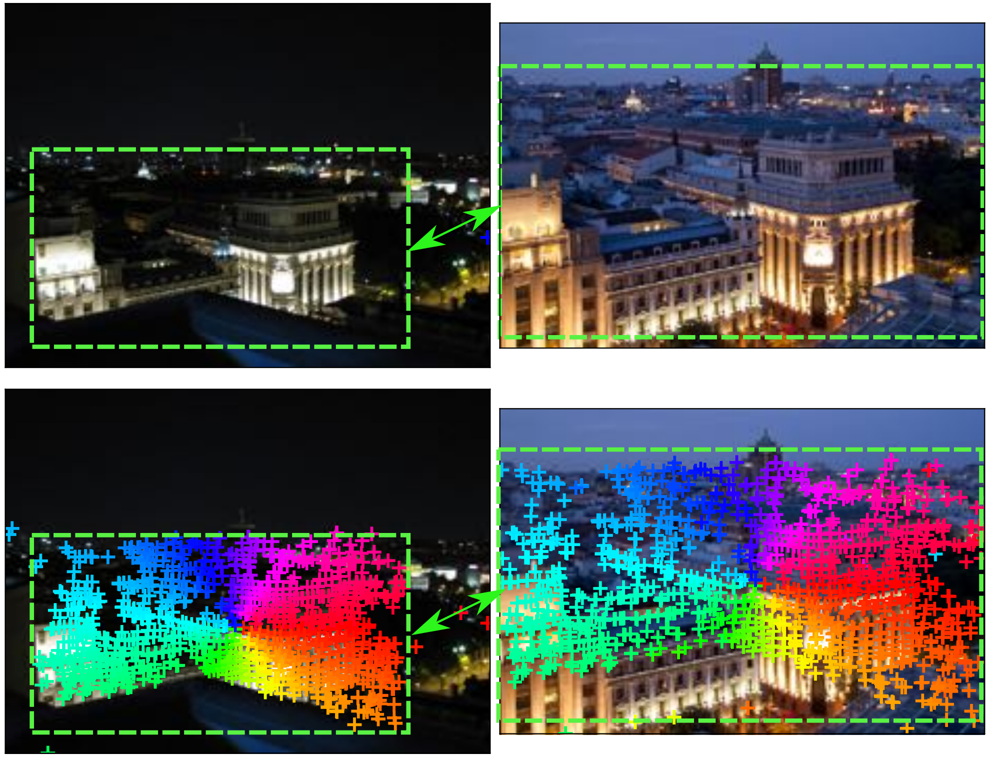
Image with
top-scored keypoints

repeatability map

reliability map



Example of Feature Matching using R2D2



The colored crosses indicate matched keypoints. As can be seen, our method even works under very challenging conditions such as day-night image pairs and large view point changes.

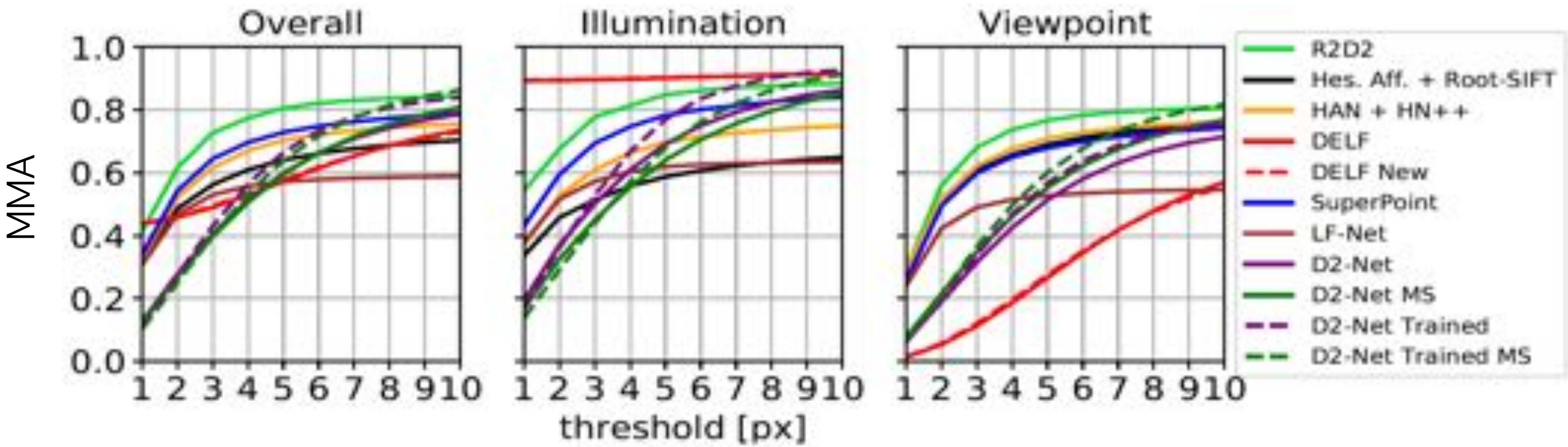
Results

HPatches

116 sequences of 6 images

- 57 containing large changes in illumination
- 59 containing large changes in viewpoint

Results



- R2D2 outperforms the state of the art on HPatches.
- The metric used is Mean Matching Accuracy (MMA).

Detailed Results on the Aachen Day-Night Dataset

DEVIEW 2019

	method	accuracy	#kpts	Dim	#weights
Classic approach	RootSIFT[25]	65.3	11K	128	-
	HAN+HN[30]	75.5	11K	128	2M
Magic Leap	SuperPoint[9]	75.5	7K	256	1.3M
Google	DELF (new)[32]	85.7	11K	1024	9M
Benchmark creators	D2-Net[11]	**88.8**	19K	512	15M
	R2D2	**88.8**	10K	128	1.0M

- Accuracy (higher is better)
  ➔ R2D2: outperforms all other approaches, including recent ones
- Number of keypoints (less is better)
  ➔ R2D2: equal or less than other approaches
- Feature dimension (less is better)
  ➔ R2D2: much smaller than other top-ranking approaches (up to 8x smaller)
- Model size (memory, less is better)
  ➔ R2D2: much smaller than other top-ranking approaches (up to 15x smaller)

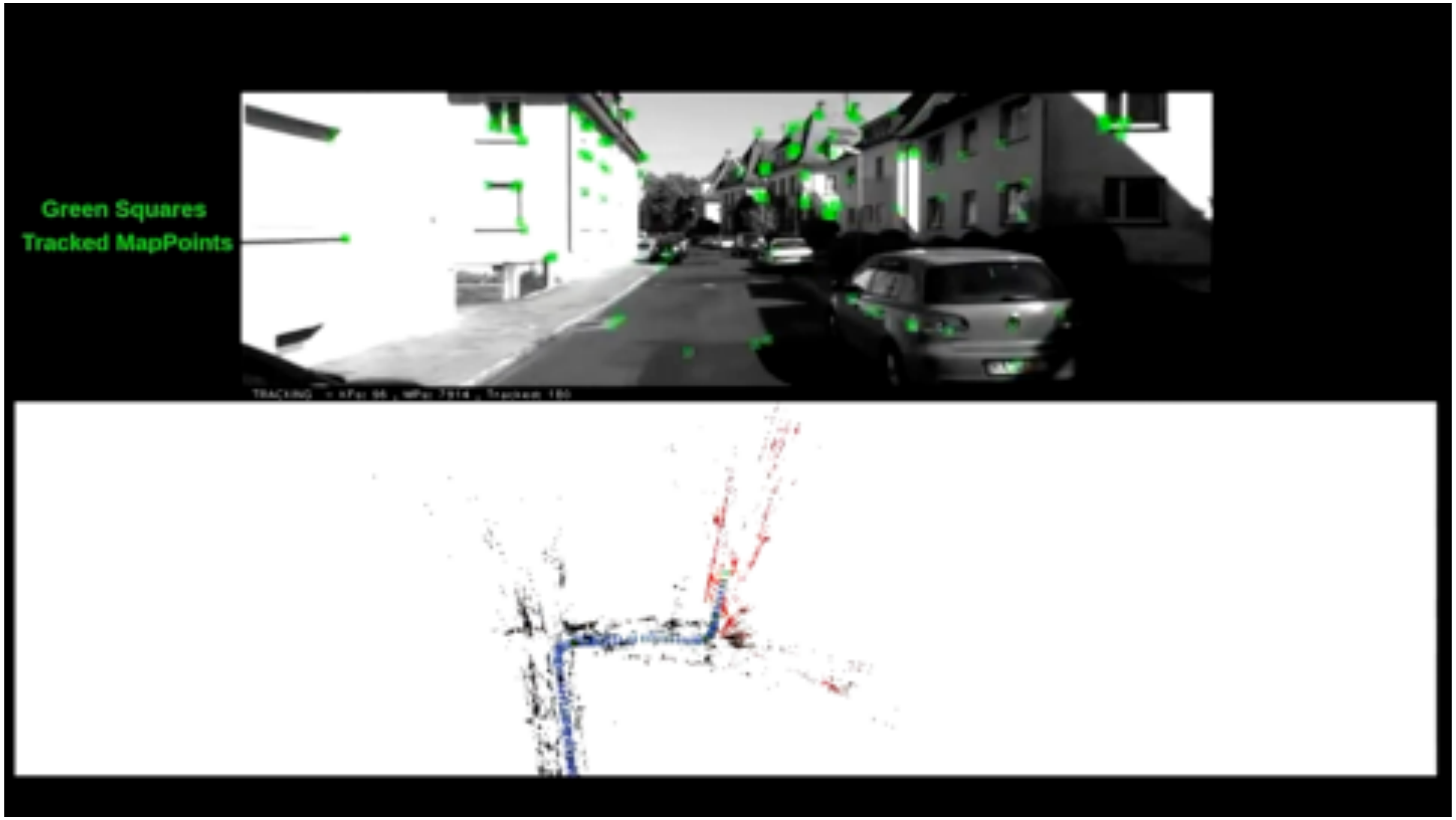
Code and models will be released!

VSLAM in Dynamic Environments

DEVIEW 2019

VSLAM – Visual Simultaneous Localization and Mapping

Example: ORB-SLAM2

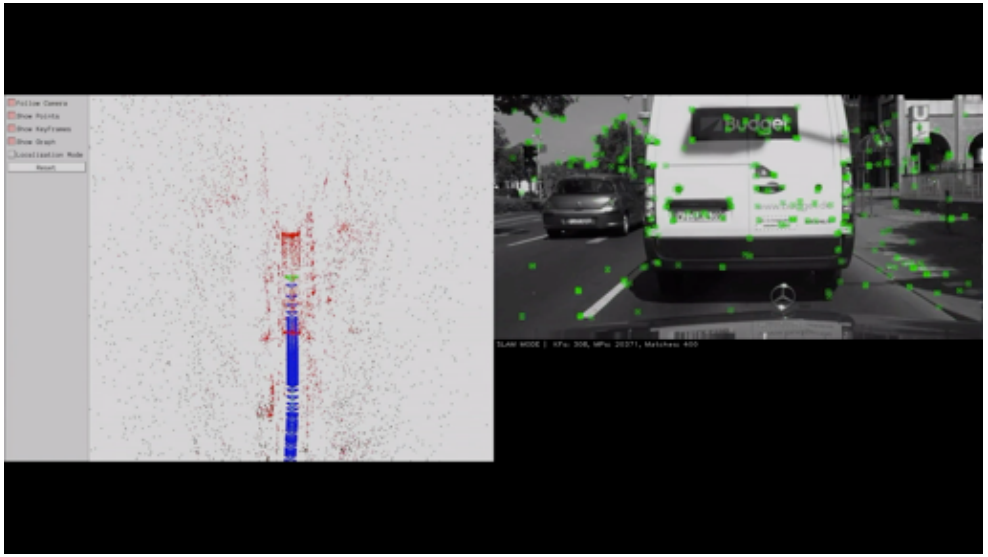


Raúl Mur-Artal and Juan D. Tardós, ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras

VSLAM in Dynamic Environments

Challenge:

- Structure-based VSLAM assumes that the world is static.
- Dynamic areas are treated as outliers.
- This does not work if the dynamic areas are dominant.

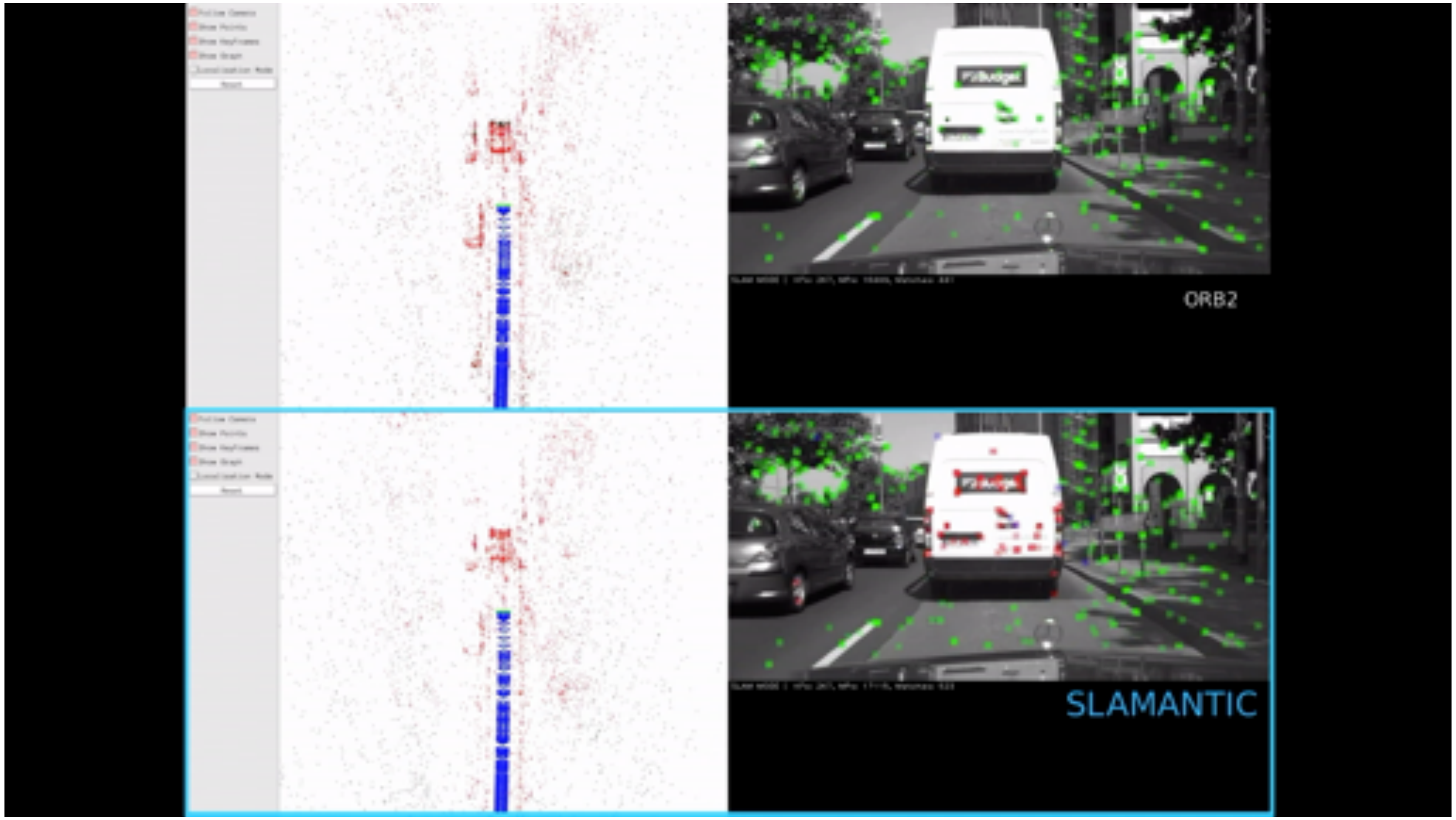


SLAMANTIC (ICCV19 workshop paper)

- We propose to use semantic information (in addition to geometry) to handle dynamic areas in the scene.
- Our approach estimates a confidence value which is used to select keypoints for the mapping part.

SLAMANTIC - Leveraging Semantics to Improve VSLAM In Dynamic Environments

DEVIEW 2019



Code is available online!

DEVIEW 2019

Conclusion

Conclusion

1. Visual Localization is an enabling technology for many applications, e.g., in robotics.
2. It is very challenging due to the ever changing world.
3. There is very good progress in the field but it is far from being solved.
4. Data driven methods might help making it more robust in the future.

Thank You

Resources

DEVIEW 2019

R2D2: https://github.com/naver/r2d2

SLAMANTIC: https://github.com/mthz/slamantic

VKITTI: https://europe.naverlabs.com/research/computer-vision/proxy-virtual-worlds/

Virtual Gallery: https://europe.naverlabs.com/research/3d-vision/virtual-gallery-dataset/

Local Features Survey: https://arxiv.org/abs/1807.10254

COLMAP: https://colmap.github.io/

OpenMVG: https://github.com/openMVG/openMVG

Visual Localization Benchmark: http://visuallocalization.net

Visual Localization Tutorial: https://sites.google.com/view/lsvpr2019/home

Baidu IBL dataset: https://sites.google.com/site/xunsunhomepage/