Learning Feature Representations for Localization and Mapping

Mihai Dusmanu
Johannes Schönberger, Marc Pollefeys
Ignacio Rocco, Tomas Pajdla, Josef Sivic, Akihiko Torii, Torsten Sattler
1. An introduction to local features

2. D2-Net – detect-and-describe approach to local features

3. Open research question

4. Multi-view keypoint refinement for accurate reconstructions
Why do we need local features?

SfM, SLAM, Visual Localization, AR…

Efficiency / scalability
What do we want from local features?

Repeatability and matchability

Robustness (viewpoint / seasonal / day-night changes, motion blur)
Detect-Then-Describe
Classic Approach

- Detectors:
  - Handcrafted: DoG, Harris, Hessian, …
  - Trainable: TILDE, TCDET, Quad-Networks, …
  - Hybrid: HesAffNet

- Descriptors:
  - SIFT, BRIEF, …
  - T-Feat, HardNet, GeoDesc, …

- Full pipeline:
  - SIFT, ORB, …
  - LIFT, LF-Net, …
Describe-Then-Detect
DELF
Large-Scale Image Retrieval with Attentive Deep Local Features
Noh et al., ICCV 2017

Shared Encoder
SuperPoint
SuperPoint: Self-Supervised Interest Point Detection and Description
DeTone et al., CVPR Workshops 2018
D2-Net – Detect-and-Describe

VGG, ResNet, MobileNet, …

feature extraction

local descriptor $d_{ij}$
D2-Net – What layer to choose?

Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, ECCV 2014

Low-Level Features

Mid-Level Features

High-Level Features
D2-Net – What layer to choose? Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, ECCV 2014

Low-Level Features

Mid-Level Features

High-Level Features

Low-Level Features

Mid-Level Features

High-Level Features

VGG 16

conv 3x3, 64

ReLu

maxpool 2x2, stride 2

conv 3x3, 128

conv 3x3, 128

maxpool 2x2, stride 2

conv 3x3, 256

conv 3x3, 256

maxpool 2x2, stride 2

conv 3x3, 512

conv 3x3, 512

conv 3x3, 512

conv 3x3, 512

maxpool 2x2, stride 2
D2-Net – What layer to choose? Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, ECCV 2014

Low-Level Features

Mid-Level Features

High-Level Features

VGG 16

conv 3x3, 64
maxpool 2x2, stride 2
conv 3x3, 128
conv 3x3, 128
maxpool 2x2, stride 2
conv 3x3, 256
conv 3x3, 256
maxpool 2x2, stride 2
conv 3x3, 512
conv 3x3, 512
conv 3x3, 512
maxpool 2x2, stride 2
\((i, j)\) is a detection \iff \(D_{ij}^k\) is a local max. in \(D^k\),
with \(k = \arg \max_t D_{ij}^t\).
(i, j) is a detection $\iff D^k_{ij}$ is a local max. in $D^k_i$, with $k = \arg \max_t D^t_{ij}$. 
(i, j) is a detection $\iff D_{ij}^k$ is a local max. in $D^k$, with $k = \arg \max_t D_{ij}^t$. 
D2-Net – Soft Keypoint Detection for Training

\[
\beta_{ij}^k = \frac{D_{ij}^k}{\max_t D_{ij}^t}, \quad \alpha_{ij}^k = \frac{\exp(D_{ij}^k)}{\sum_{(i',j') \in \mathcal{N}(i,j)} \exp(D_{i'j'}^k)}
\]

\[
s_{ij} \propto \max_k \left( \alpha_{ij}^k \beta_{ij}^k \right)
\]
D2-Net – Joint Detection-Description Loss

- Triplet loss for description

\[ m(c) = \max(0, M + p(c)^2 - n(c)^2) \]

  - Negative sample: in-image-pair negative mining - filter out repetitive structures

- Weighted average of triplet losses over all correspondences

\[ \mathcal{L}(I_1, I_2) = \sum_{c \in C} \left( \frac{s_c^{(1)} s_c^{(2)}}{\sum_{q \in C} s_q^{(1)} s_q^{(2)}} \right) m(p(c), n(c)) \]

  - good correspondence \(\iff\) low triplet loss value

- Requires correspondences

  - MegaDepth: 196 different locations reconstructed with COLMAP SfM / MVS

MegaDepth: Learning Single-View Depth Prediction from Internet Photos, Li & Snavely, CVPR 2018
D2-Net – Results

- Long-term Visual Localization Benchmark
  - https://www.visuallocalization.net
  - Different localization scenarios:
    - Different seasons / illumination conditions (including night-to-day)
    - Indoor localization
    - Autonomous driving
    - Suburban / Park scenes with vegetation
  - Ranked #1 on 3 datasets and #2 on 2 datasets
  - Using NetVLAD / VLAD retrieval
D2-Net – Summary

- Joint detection and description
- State-of-the-Art for local features on challenging camera localization tasks
- Versatile: not architecture-specific
- Problems:
  - Poor keypoint localization
    - Raw matching: beats SOTA starting at ~6px
    - >1 px reprojection error for 3D reconstructions
    - Large receptive field, max pooling
  - Feature ambiguity
    - Large receptive field
Will there be a Swiss-army-knife local feature?
An extreme example…

- High level detections: robust but not well localized
- Low level detections: very well localized, but not as robust
Feature extraction

Independent for each image
Feature extraction

Independent for each image

tentative matches
Feature extraction

Independent for each image

Patch matching

Patch Match CNN

tentative matches

local geometric transformation
Refining with multiple views
Tentative matches graph

Challenges:
- Incorrect matches
- Inaccurate transformations
- (Very) large graphs
- Feature drift
- Repeated structures

\[
\min_{x_k} \sum_{i \to j \text{ edge}} \rho \left( \| x_j - T_{ij} x_i \|^2 \right)
\]
Questions