Detect-to-Retrieve: Efficient Regional Aggregation for Image Search

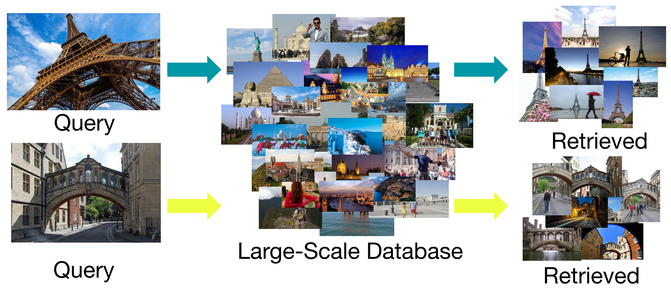
Marvin Teichmann*[1], André Araujo*[2], Menglong Zhu[2], Jack Sim[2]

Code+data: github.com/tensorflow/models/tree/master/research/delf

[1]University of Cambridge, UK [2]Google AI

*Equal contribution

Large-Scale Image Retrieval

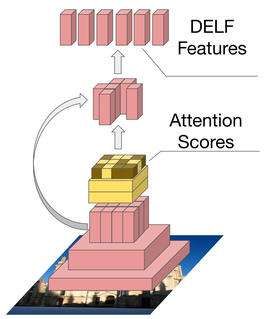


Challenges	Contributions
Low performance on small objects	Regional representation using object detector
Indexing multiple regions is inefficient	Regional Aggregated Matching Kernels
No manually labeled landmark box datasets	Google Landmark Boxes dataset

State-of-the-art results on Revisited Oxford/Paris datasets

Detect-to-Retrieve

Background: DELF



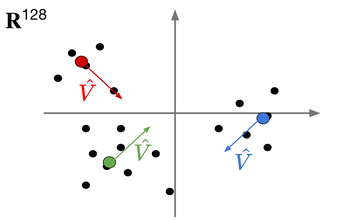
Regional Search

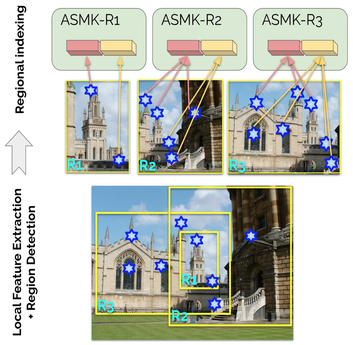
- ✓ DELF-ASMK extracted for each box
- ✓ Each box indexed independently
- ✓ Index size = O(#boxes in all images)
- ✓ Retrieval system:
  - score for an image = its maximum box score

Background: ASMK

$$\text{sim}^{\text{ASMK}} = \sum_c \text{ASMK}(\mathcal{X}_c, \mathcal{Y}_c)$$

$$\text{ASMK}(\mathcal{X}_c, \mathcal{Y}_c) = \sigma\left(\hat{V}(\mathcal{X}_c)^\top \hat{V}(\mathcal{Y}_c)\right)$$





Regional Aggregation: R-ASMK

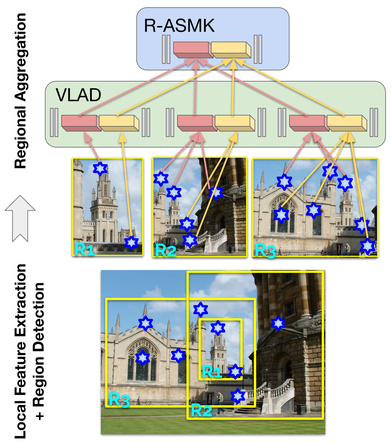
- ✓ Extension of aggregated matching kernels
  - Pooling over boxes, then images
  - Boxes drive improved representation
- ✓ Motivation:
  - average pooling of VLAD similarities
  - + renormalization
  - + selectivity function
- ✓ Index size = O(#images)
  → Memory-efficient, especially important for large-scale setting

$$\text{sim}^{\text{R-ASMK}} = \sum_c \text{R-ASMK}(\{\mathcal{X}_c^r\}_r, \{\mathcal{Y}_c^r\}_r)$$

$$\text{R-ASMK}(\{\mathcal{X}_c^r\}_r, \{\mathcal{Y}_c^r\}_r) = \sigma\left(\hat{V}_R(\{\mathcal{X}_c^r\}_r)^\top \hat{V}_R(\{\mathcal{Y}_c^r\}_r)\right)$$

$$\hat{V}_R(\{\mathcal{Y}_c^r\}_r) = \frac{V_R(\{\mathcal{Y}_c^r\}_r)}{\|V_R(\{\mathcal{Y}_c^r\}_r)\|}$$

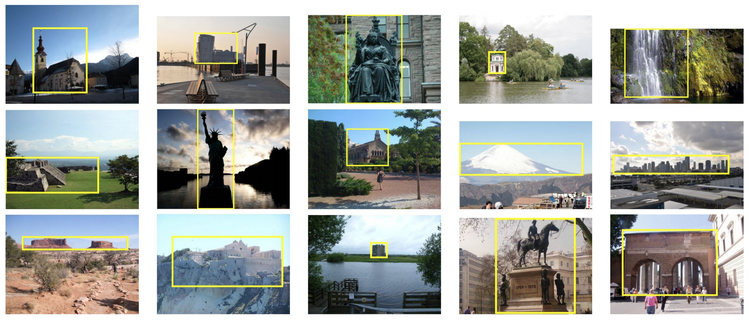
$$V_R(\{\mathcal{Y}_c^r\}_r) = \frac{1}{R_n} \sum_r \gamma(\mathcal{Y}^r) V(\mathcal{Y}_c^r)$$

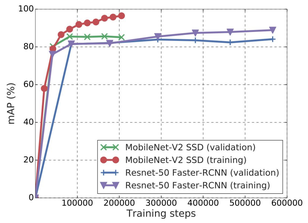


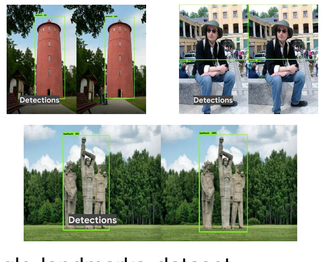
- ✓ Alternative naive regional kernels
  - E.g., R-VLAD: average pooling of VLAD's
  - Normalization issues: per-VW information is "watered down"

Google Landmark Boxes

- ✓ 86k annotated boxes, from 15k landmarks
- ✓ One box per image capturing most prominent landmark
- ✓ Accurate detection with off-the-shelf architectures



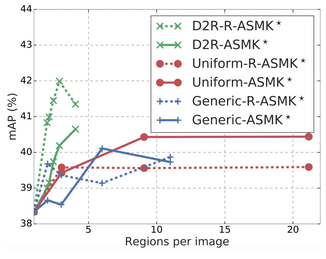




https://www.kaggle.com/google/google-landmarks-dataset

Experimental Results

Ablation Study



- ✓ D2R improves upon no-detection baseline
  - 2.31% for regional search
  - 3.65% for regional aggregation
- ✓ Regional aggregation > Regional Search
  - Higher mAP, smaller index
- ✓ D2R > uniform, generic detectors

State-of-the-Art Image Retrieval Results

- ✓ Best results in Revisited datasets for all protocols and metrics
- ✓ 9.3% mAP improvement on Revisited Oxford (Hard)
- ✓ 1.9% mAP improvement on Revisited Paris (Hard)

Method	Medium								Hard							
	$\mathcal{R}$Oxf		$\mathcal{R}$Oxf+$\mathcal{R}$1M		$\mathcal{R}$Par		$\mathcal{R}$Par+$\mathcal{R}$1M		$\mathcal{R}$Oxf		$\mathcal{R}$Oxf+$\mathcal{R}$1M		$\mathcal{R}$Par		$\mathcal{R}$Par+$\mathcal{R}$1M	
	mAP	mP@10	mAP	mP@10	mAP	mP@10	mAP	mP@10	mAP	mP@10	mAP	mP@10	mAP	mP@10	mAP	mP@10
AlexNet-GeM [30]	43.3	62.1	24.2	42.8	58.0	91.6	29.9	84.6	17.1	26.2	9.4	11.9	29.7	67.6	8.4	39.6
VGG16-GeM [30]	61.9	82.7	42.6	68.1	69.3	97.9	45.4	94.1	33.7	51.0	19.0	29.4	44.3	83.7	19.1	64.9
ResNet101-R-MAC [9]	60.9	78.1	39.3	62.1	78.9	96.9	54.8	93.9	32.4	50.0	12.5	24.9	59.4	86.1	28.0	70.0
ResNet101-GeM [30]	64.7	84.7	45.2	71.7	77.2	98.1	52.3	95.3	38.5	53.0	19.9	34.9	56.3	89.1	24.7	73.3
ResNet101-GeM†+DSM [34]	65.3	87.1	47.6	76.4	77.4	99.1	52.8	96.7	39.2	55.3	23.2	37.9	56.2	89.9	25.0	74.6
HesAff-rSIFT-ASMK* [38]	60.4	85.6	45.0	76.0	61.2	97.9	42.0	95.3	36.4	56.7	25.7	42.1	34.5	80.6	16.5	63.4
HesAff-rSIFT-ASMK*+SP [38]	60.6	86.1	46.8	79.6	61.4	97.9	42.3	95.3	36.7	57.0	26.9	45.3	35.0	81.7	16.8	65.3
HesAff-HardNet-ASMK*+SP [24]	65.6	90.2	–	–	65.2	98.9	–	–	41.1	59.7	–	–	38.5	87.9	–	–
DELF-ASMK*+SP [25, 28]	67.8	87.9	53.8	81.1	76.9	99.3	57.3	98.3	43.1	62.4	31.2	50.7	55.4	**93.4**	26.4	75.7
DELF-ASMK* (reimpl.)	65.7	87.9	–	–	77.1	98.7	–	–	41.0	57.9	–	–	54.6	90.9	–	–
DELF-D2R-R-ASMK* (ours)	69.9	89.0	–	–	78.7	99.0	–	–	45.6	61.9	–	–	57.7	93.0	–	–
— DELF-GLD (ours)	73.3	90.0	61.0	84.6	**80.7**	99.1	**60.2**	97.9	47.6	64.3	33.6	53.7	**61.3**	**93.4**	**29.9**	82.4
DELF-ASMK*+SP (reimpl.)	68.9	90.9	–	–	76.6	98.7	–	–	46.6	66.7	–	–	52.2	87.6	–	–
DELF-D2R-R-ASMK*+SP (ours)	71.9	91.3	–	–	78.0	**99.4**	–	–	48.5	66.7	–	–	54.0	87.6	–	–
— DELF-GLD (ours)	**76.0**	**93.4**	**64.0**	**87.7**	80.2	99.1	59.7	**99.0**	**52.4**	**70.9**	**38.1**	**61.3**	58.6	91.0	29.4	**83.9**

Qualitative Results