Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking

Filip Radenović[1] Ahmet Iscen[1] Giorgos Tolias[1] Yannis Avrithis[2] Ondřej Chum[1]

[1]Visual Recognition Group, Czech Technical university in Prague [2]INRIA, Rennes

Oxford 5k and Paris 6k
What was wrong with our favorite datasets?

- **Annotation errors:** skewed comparison of different methods

Original labeling mistakes: **Query (blue)** image and the associated database images that were originally marked as **negative (red)** or **positive (green)**.

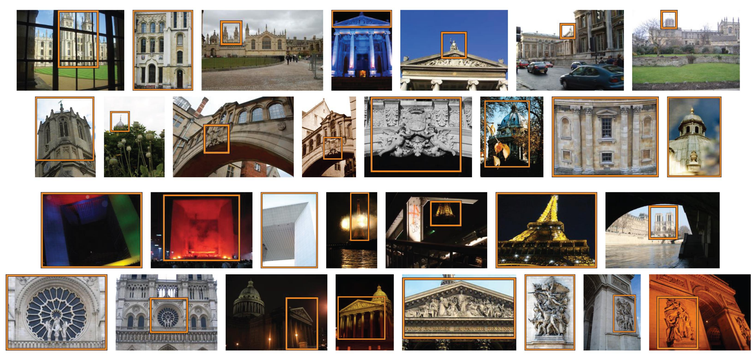
- **Solved:** saturated performance, every challenging image labeled as *Junk*
- **Over-fitting:** small datasets, extension Oxford 100k (easy, false negatives)

Examples of false negative images in Oxford100k.

What's New

- Errors in the annotation are fixed
- *Labeling of all* images is revisited
- New distractor dataset with 1 million images is created
- Images are chosen to be challenging for these two benchmarks
- New set of 15 queries per benchmark is added
- New set of evaluation protocols with increasing difficulty: Easy (E), Medium (M), and Hard (H)

New Queries



Oxford (first two rows), and Paris (second two rows)

Revisiting Annotation and Evaluation

- **Annotation procedure:**
  - **Step1:** Selection of potential positives
  - **Step2:** Label assignment, ***Easy***, ***Hard***, ***Unclear***, and ***Negative***
  - **Step3:** Refinement voting for consensus among 5 annotators
- **Instructions to annotators:**
  - ***Easy:*** Clearly depicts same side (or symmetry), no significant change
  - ***Hard:*** Same side (or symmetry), difficult viewing conditions
  - ***Unclear:*** Context has to be used to make a decision, different side but partially symmetric with the query side
  - ***Negative:*** None of the previous conditions satisfied

Query (blue) images and images that are respectively marked as **easy (dark green)**, **hard (light green)**, and **unclear (yellow)**.

- **Three new evaluation setups:**
  - **Easy**: Positive = Easy images, Ignore = Hard & Unclear images
  - **Medium**: Positive = Easy & Hard images, Ignore = Unclear images
  - **Hard**: Positive = Hard images, Ignore = Easy & Unclear images

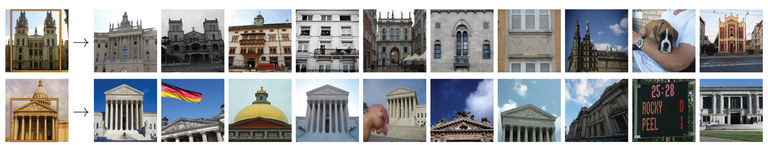
$\mathcal{R}$Oxford				
Labels	Easy	Hard	Uncl.	Neg.
Positive	438	50	93	1
Junk	50	222	72	9
Negative	1	72	133	63768

$\mathcal{R}$Paris				
Labels	Easy	Hard	Uncl.	Neg.
Positive	1222	643	136	6
Junk	91	813	835	61
Negative	16	147	273	71621

Number of images according to label swap from original annotation (positive, junk, negative) to the new one (easy, hard, unclear, negative)

Distractor set of 1M images

- New distractors set of 1,001,001 high-resolution (1024 x 768) images
- Significantly more challenging than Oxford100k, in size and difficulty
- Made to be more distracting by combining state-of-the-art methods



The most distracting images per query for two queries.

Extensive evaluation

mAP Old vs New

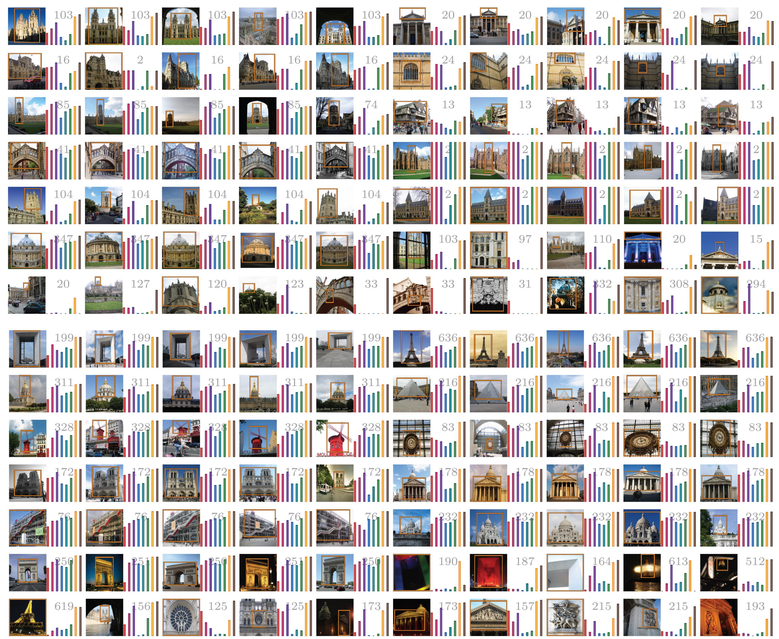
Method	Oxf	$\mathcal{R}$Oxford			Par	$\mathcal{R}$Paris		
		E	M	H		E	M	H
HesAff–rSIFT–SMK*	78.1	74.1	59.4	35.4	74.6	80.6	59.0	31.2
R–[O]–R-MAC	78.3	74.2	49.8	18.5	90.9	89.9	74.0	52.1
R–[FT]–GeM	87.8	84.8	64.7	38.5	92.7	92.1	77.2	56.3
R–[FT]–GeM+DFS	90.0	86.5	69.8	40.5	95.3	93.9	88.9	78.5

Time and Memory

Method	Memory (GB)	Time (sec)		
		Extraction		Search
		GPU	CPU	
HesAff–rSIFT–ASMK*	62.0	n/a + 0.06	1.08 + 2.35	0.98
HesAff–rSIFT–ASMK*+SP				2.00
DELF–ASMK*+SP	10.3	0.41 + 0.01	n/a + 0.54	0.52
A–[FT]–GeM	0.96	0.12	1.99	0.38
V–[FT]–GeM	1.92	0.23	31.11	0.56
R–[FT]–GeM	7.68	0.37	14.51	1.21

State-of-the-art performance

Method	Medium				Hard			
	$\mathcal{R}$Oxf+$\mathcal{R}$1M		$\mathcal{R}$Par+$\mathcal{R}$1M		$\mathcal{R}$Oxf+$\mathcal{R}$1M		$\mathcal{R}$Par+$\mathcal{R}$1M	
	mAP	mP@10	mAP	mP@10	mAP	mP@10	mAP	mP@10
HesAff–rSIFT–VLAD	17.4	34.8	19.6	76.1	5.6	7.0	3.3	21.1
HesAff–rSIFT–SMK*+SP	38.1	67.1	34.5	89.3	17.7	30.3	11.0	49.1
HesAff–rSIFT–ASMK*+SP	46.8	79.6	42.3	95.3	26.9	45.3	16.8	65.3
DELF–ASMK*+SP	53.8	81.1	57.3	98.3	31.2	50.7	26.4	75.7
R – [O] – MAC	24.2	43.7	40.8	93.0	5.7	14.4	18.2	67.7
R – [O] – SPoC	21.5	40.4	41.6	92.0	2.8	5.6	15.3	54.4
R – [O] – CroW	21.2	39.4	42.7	92.9	3.3	9.3	16.3	61.6
R – [O] – GeM	25.6	45.1	46.2	94.0	4.7	13.4	20.3	70.4
R – [O] – R-MAC	29.2	48.9	49.3	93.7	4.5	13.0	21.3	67.4
R – [FT] – GeM	45.2	71.7	52.3	95.3	19.9	34.9	24.7	73.3
R – [FT] – R-MAC	39.3	62.1	54.8	93.9	12.5	24.9	28.0	70.0
Query expansion (QE) and diffusion (DFS)								
HesAff–rSIFT–HQE	42.7	67.4	44.2	90.1	23.2	37.6	20.3	51.4
HesAff–rSIFT–HQE+SP	52.0	76.7	46.8	93.0	29.8	50.1	21.8	61.9
DELF–HQE+SP	60.6	79.7	65.2	96.1	37.9	56.1	35.8	69.1
R – [FT] – GeM+αQE	49.0	74.7	58.0	95.9	24.2	40.3	31.0	80.4
R – [FT] – GeM+DFS	61.5	77.1	84.9	95.9	33.1	48.2	71.6	93.7
R – [FT] – R-MAC+DFS	56.6	68.6	83.2	93.3	28.4	43.6	70.4	89.1
HesAff–rSIFT–ASMK*+SP → R–[FT]–GeM+DFS	74.3	87.9	85.9	97.1	48.7	65.9	73.2	96.6
HesAff–rSIFT–ASMK*+SP → R–[FT]–R-MAC+DFS	74.9	87.9	87.5	97.1	47.5	62.4	76.0	96.3
DELF–ASMK*+SP → R–[FT]–R-MAC+DFS	68.7	83.6	86.6	98.1	39.4	55.7	74.2	94.6



Methods: **HessAff–rSIFT–ASMK*+SP**, **DELF–ASMK*+SP**, **DELF–HQE+SP**, **V–[O]–R-MAC**, **R–[O]–GeM**, **R–[FT]–GeM**, **R–[FT]–GeM+DFS**, **HessAff–rSIFT–ASMK*+SP->R–[FT]–GeM+DFS**