Multiple Measurements and Joint Dimensionality Reduction for Large Scale Image Search with Short Vectors

Filip Radenović[1] Hervé Jégou[2] Ondřej Chum[1]

[1] Centre for Machine Perception, CTU in Prague

[2] INRIA, Rennes

Short-vector image retrieval with multiple vocabularies

Query

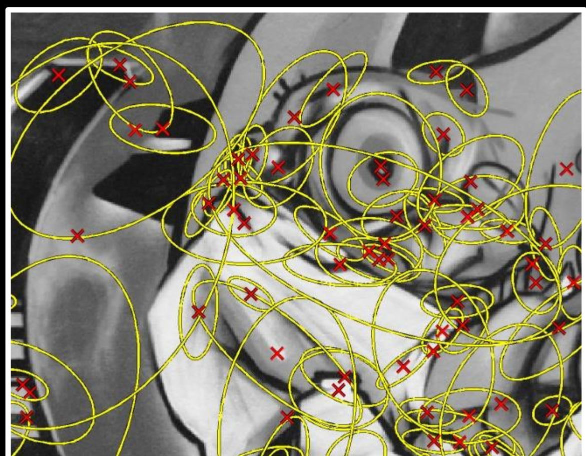
Results

Small memory footprint of the dataset,
each image represented by a short vector (128D)

Our approach is based on bag-of-words (BOW)
multiple vocabularies (multiple BOW) are used
to reduce quantization effect

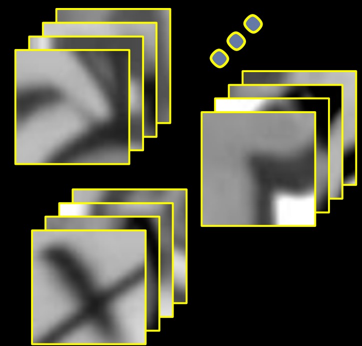
Bag-of-words (BOW) baseline

Keypoint Detection

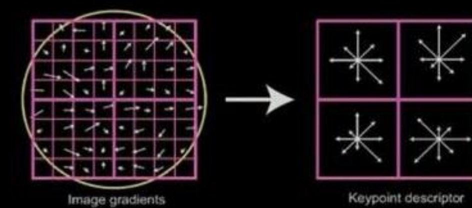


graffiti

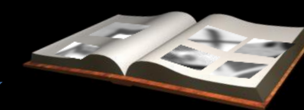
Local Appearance



SIFT Description [Lowe – IJCV 2004]



Visual Vocabulary



word_1, word_2, word_2, word_2, word_8, …, word_{100000}, word_{100002}, word_{100002}, …, word_{948534}, word_{998125}, …

graffiti visual words

L2 normalized histogram of occurrences – BOW vector

*BOW vectors compared using cosine similarity

Sivic & Zisserman – ICCV 2003
Video Google: A Text Retrieval Approach to Object Matching in Videos

PCA dimensionality reduction and whitening

High dimensional sparse BOW image representation

*Search is done using inverted files

PCA →

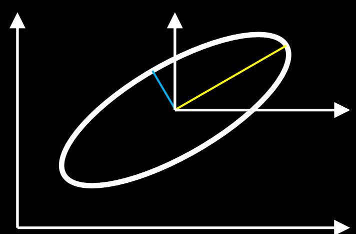
128 dimensional dense image representation

*Search is done using (approximate) nearest-neighbors

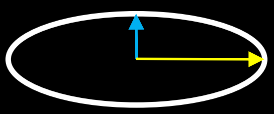
PCA dimensionality reduction and whitening

Jegou & Chum analyze effects of different parts of PCA on BOW vectors:

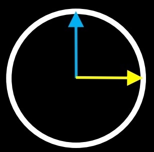
- Centering – emphasize negative evidence,
  higher importance of jointly missing visual words



- PCA rotation – decorrelating and allowing
  to remove least informative dimensions



- Whitening – addresses over-counting
  (burstiness, co-occurence)



Jegou & Chum – ECCV 2012
Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening

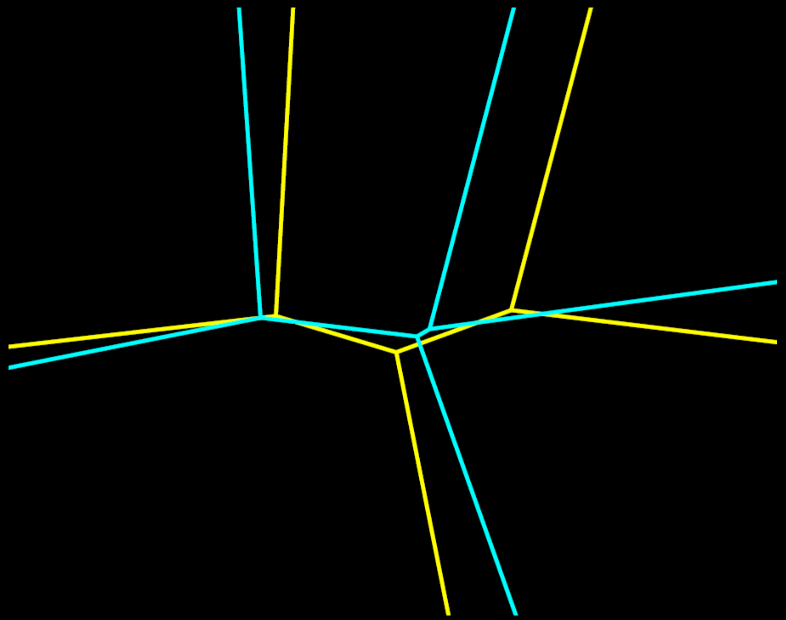
Joint dimensionality reduction of multiple vocabularies (mVocab) baseline

Joint dimensionality reduction of multiple vocabularies:

1. Multiple vocabularies are built using different k-means initializations

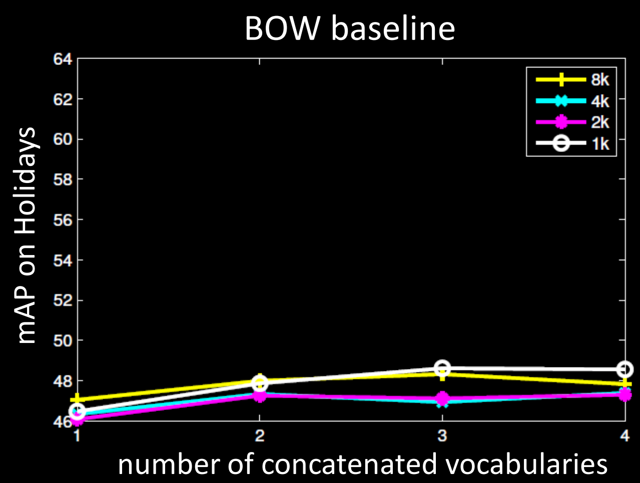
2. BOW vectors are concatenated

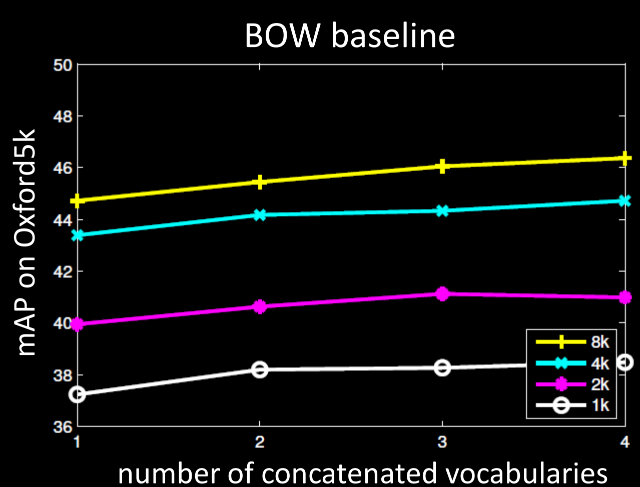
3. Concatenated BOW vectors are jointly PCA-reduced and whitened



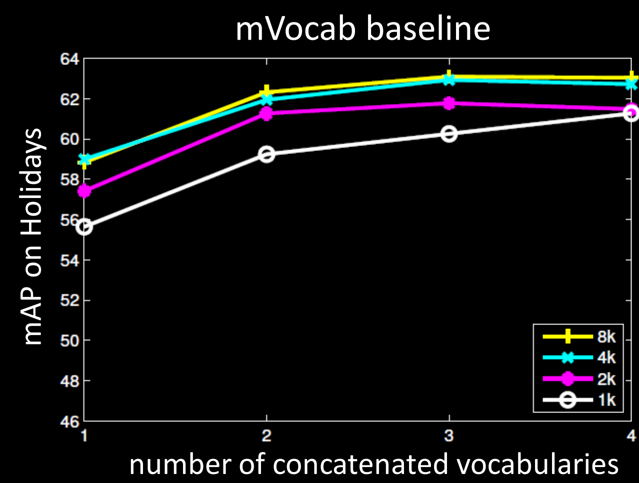
BOW vs. mVocab

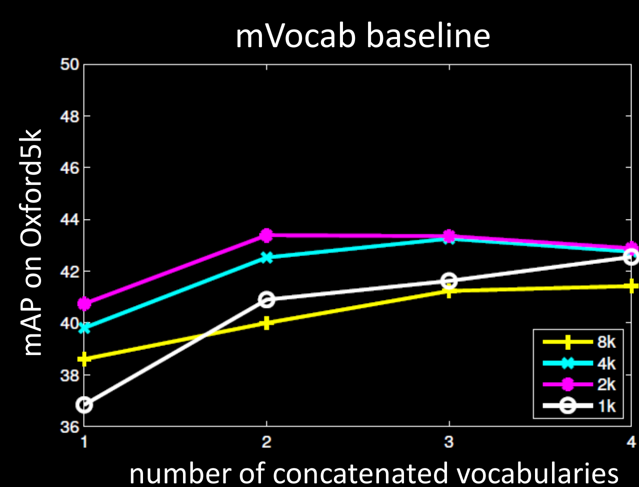
Increasing dimensionality – *n × k*





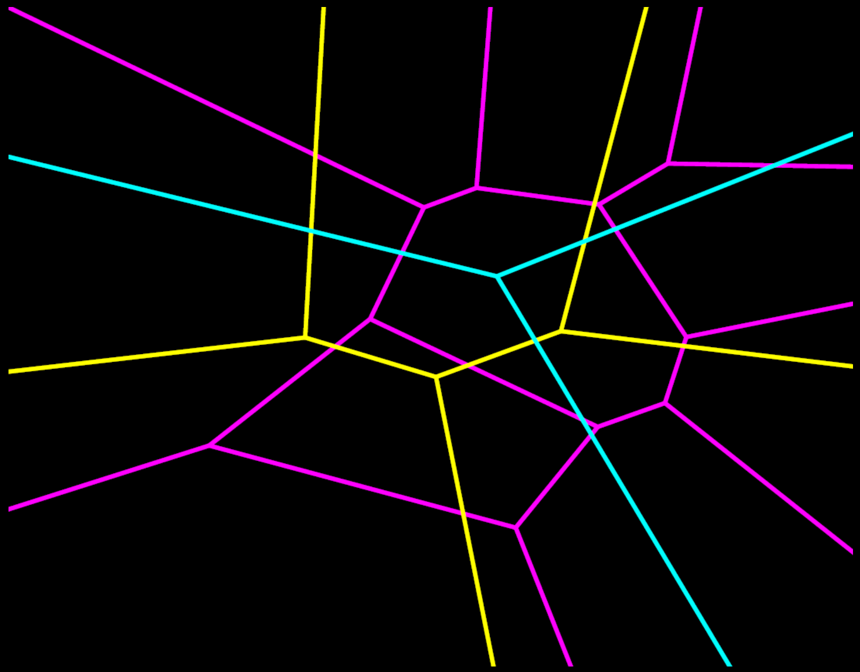
Fixed dimensionality – *128*





Multiple vocabularies with different sizes

Concatenating vocabularies with multiple sizes [Jegou & Chum – ECCV 2012], example: 4k+2k+1k+512+256+128



Grauman & Darrell – ICCV 2005
The pyramid match kernel

Stwenius & Nister – CVPR 2006
Scalable recognition with a vocabulary tree

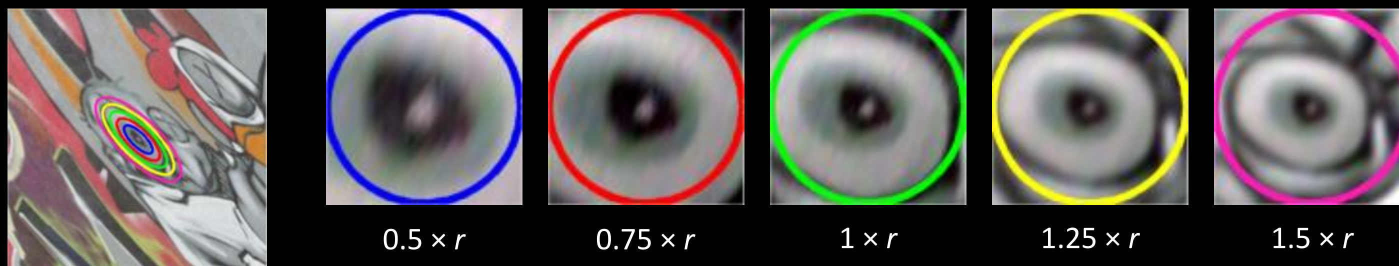
Proposed methods

Build independent (less correlated) vocabularies by:

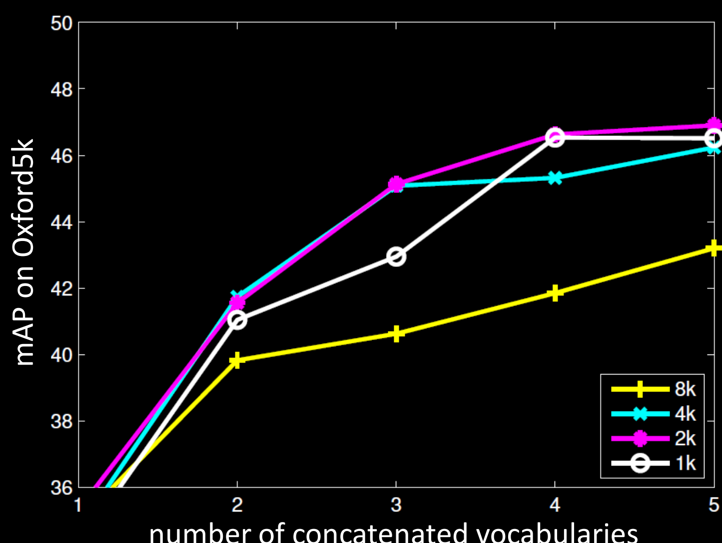
1. Using different measurement regions for calculating SIFT descriptors (mMeasReg)
   - Descriptors extracted from different image patches
2. Using different power-law normalizations of SIFT descriptors (mRootSIFT)
   - Non-linear transformations of the descriptors (and distances)
3. Using different PCA-reduced SIFT descriptors (mPCA-SIFT)
   - Linear transformation of the descriptors (and distances)

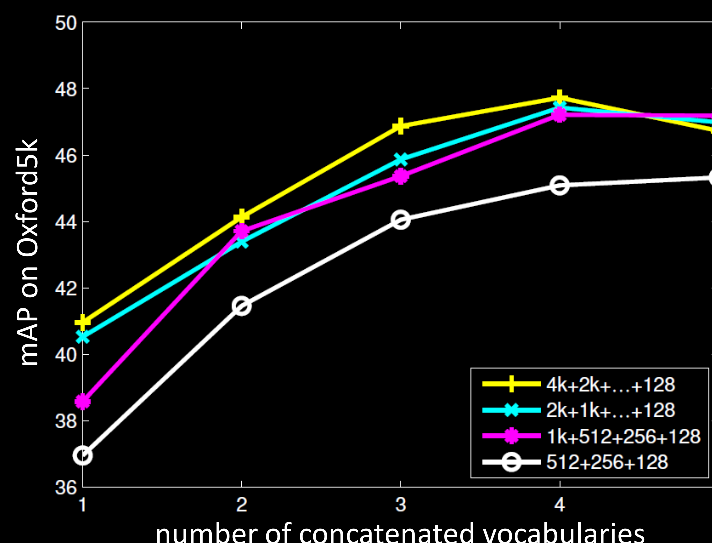
Multiple measurement regions (mMeasReg)

Construct vocabularies at multiple relative scales of the measurement regions:



r = 3√3 – relative change in the measured area radius compared to detected area radius



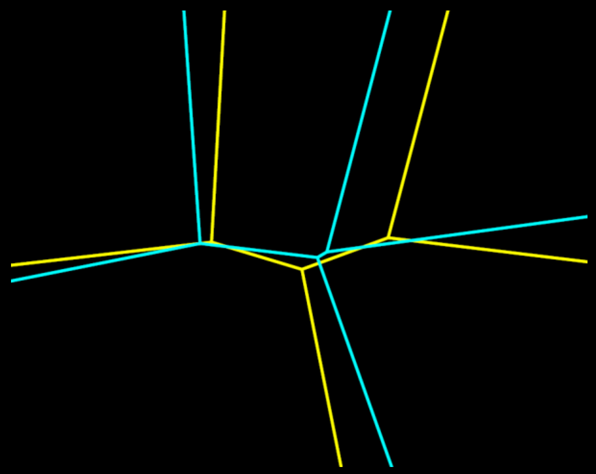


Multiple power-law normalized SIFT descriptors (mRootSIFT)

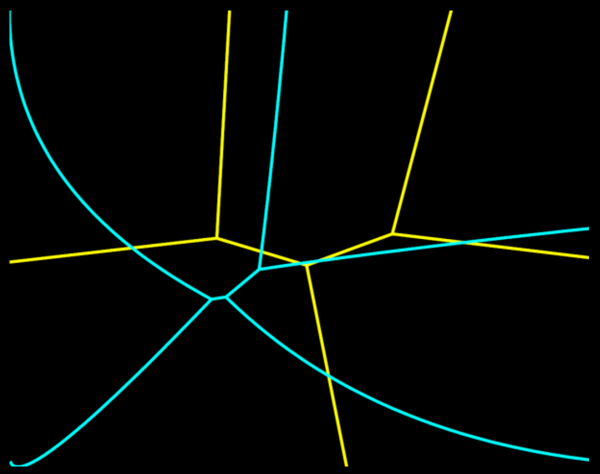
K-means with different power-law normalized SIFT descriptors result in different hypersurfaces in original SIFT descriptor space:

- SIFT descriptors + Euclidian distance = hyperplanes in SIFT space
- Rooted SIFTs + Euclidian distance = curved hypersurfaces in SIFT space

different k-means initialization

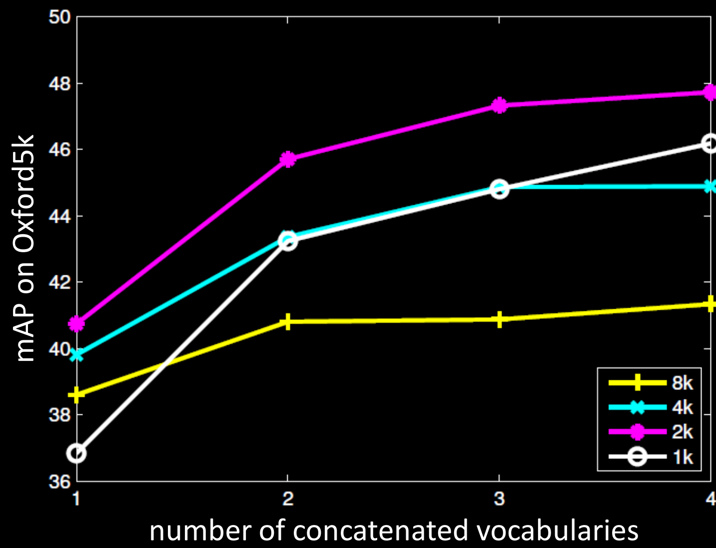


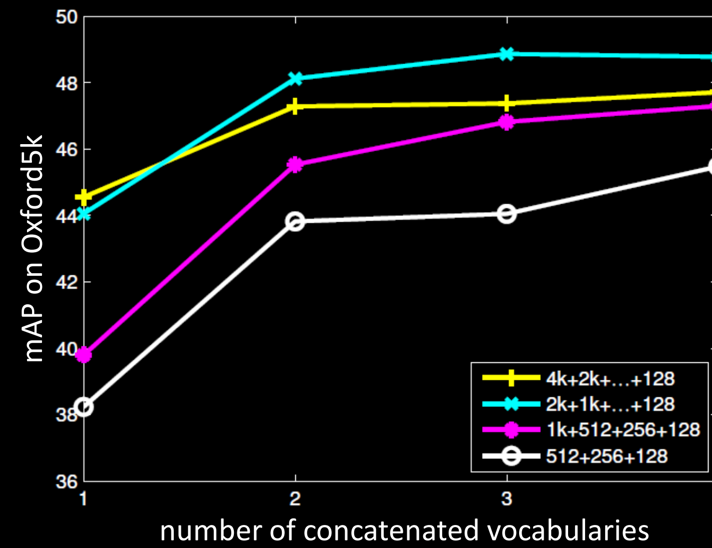
SIFT and RootSIFT



Multiple power-law normalized SIFT descriptors (mRootSIFT)

- We combine SIFT and SIFT with every component to the power of 0.4 ($SIFT^{0.4}$), 0.5 ($SIFT^{0.5}$), 0.6 ($SIFT^{0.6}$) to create four different vocabularies





Multiple linear projections of SIFT descriptors (mPCA-SIFT)

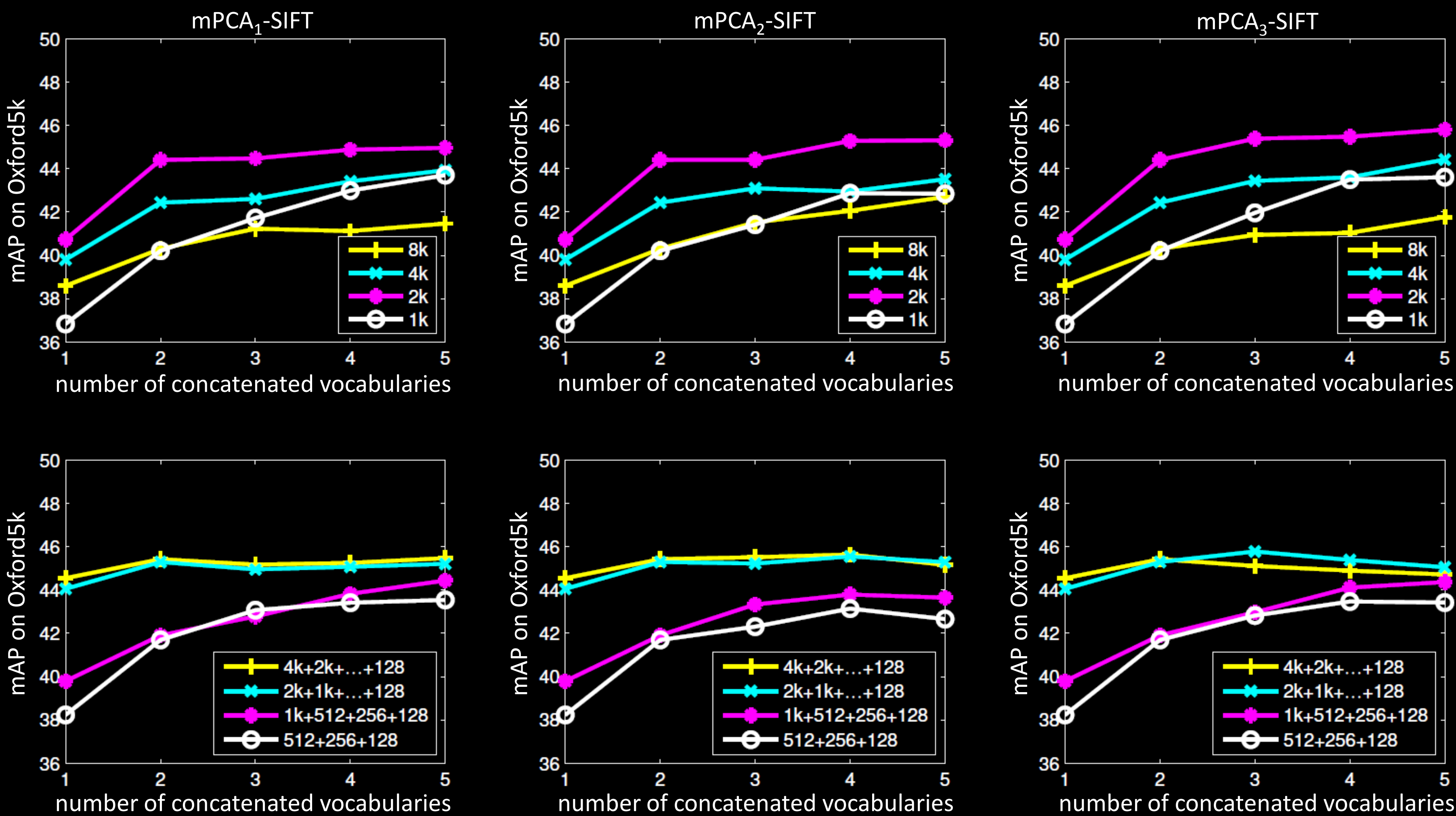
Construct vocabularies using different PCA projections of SIFTs:

1. Reduce SIFTs to 80, 64, 48, 32 dimensions for every new vocabulary while learning eigenvectors on Paris6k (mPCA$_1$-SIFT)

2. Reduce SIFTs to 80 dimensions for every new vocabulary while learning eigenvectors on different datasets: Paris6k, Holidays, UKB, PASCAL VOC'07 (mPCA$_2$-SIFT)

3. Reduce SIFTs to 80, 64, 48, 32 dimensions for every new vocabulary while learning eigenvectors on different datasets: Paris6k, Holidays, UKB, PASCAL VOC'07 (mPCA$_3$-SIFT)

Multiple linear projections of SIFT descriptors (mPCA-SIFT)



Comparison with the state-of-the-art

All presented methods have short-vector (128D) image representations:

Method	Vocabulary	Oxford5k	Oxford105k	Holidays
mVocab/BOW [1]	k=4×8k	41.3/41.4*	−/33.2*	56.7/63.0*
mVocab/BOW [1]	k=2×(32k+ . . . +128)	−/42.9*	−/35.1*	60.0/64.5*
mVocab/VLAD [1]	k=4×256	−	−	61.4
mVocab/VLAD+adapt+innorm [2]	k=4×256	44.8	37.4	62.5
$\phi_\Delta+\psi_d$+RN [3]	k=16	43.3	35.3	61.7
mMeasReg/mVocab/BOW	k=5×2k	46.9	38.9	66.9
mMeasReg/mVocab/BOW	k=4×(4k+ . . . +128)	47.7	39.2	67.3
mRootSIFT/mVocab/BOW	k=4×2k	47.7	39.8	64.3
mRootSIFT/mVocab/BOW	k=4×(2k+ . . . +128)	48.8	41.4	65.6
mPCA$_3$-SIFT/mVocab/BOW	k=5×2k	45.8	38.1	63.2
mPCA$_1$-SIFT/mVocab/BOW	k=5×(4k+ . . . +128)	45.5	37.8	64.6

[1] Jegou & Chum, Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening, ECCV 2012
[2] Arandjelovic & Zisserman, All about VLAD, CVPR 2013
[3] Jegou & Zisserman, Triangulation embedding and democratic aggregation for image search, CVPR 2014

Conclusions

+ Simple implementation
+ No speed overhead
+ Small memory requirements (128D image representation)
+ **State-of-the-art exceeded by a large margin**

- Optimal combination of vocabularies still an open problem