Descriptor Learning Using Convex Optimisation

Karen Simonyan, Andrea Vedaldi, Andrew Zisserman
{karen, vedaldi, az}@robots.ox.ac.uk
Visual Geometry Group, Department of Engineering Science, University of Oxford, UK

1. INTRODUCTION

Goal

Learn discriminative keypoint descriptors for keypoint matching and object instance retrieval

What is being learnt?

- Spatial pooling regions
- Dimensionality reduction

Contribution

- Convex large-margin formulations for
  - pooling region selection
  - dimensionality reduction
- Extension to learning under very weak supervision

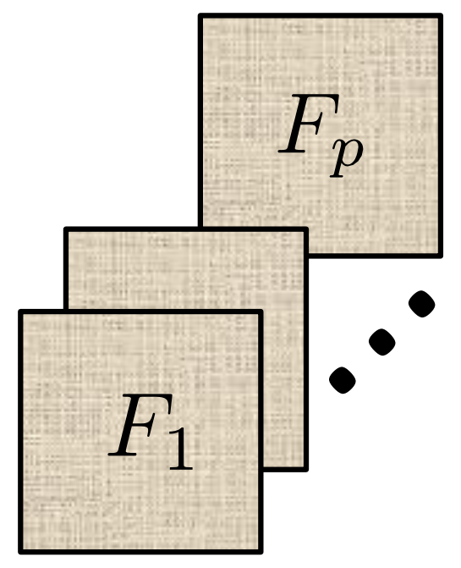
State-of-the-art in keypoint descriptor learning

- Large scale patch matching
- Large scale object retrieval

2. DESCRIPTOR COMPUTATION PIPELINE

Pre-rectified keypoint patch — E.g. scale/affine-invariant keypoint detection and rectification

Non-linear transform $\{F_c\}_{c=1}^{p}$ — SIFT-like gradient orientation binning with soft-assignment to $p=8$ feature channels



learning w *(panel 3)*

Spatial pooling $\tilde\phi = [F_c \times \Omega(w)]_{c=1}^{p}$ — Feature pooling using selected Gaussian pooling regions $\Omega(w)$ shared across feature channels

Normalisation and cropping $\phi = \min\left\{\tilde\phi/T, 1\right\}$, $T = T(F)$ — Agnostic to pooling regions configuration (useful for learning)

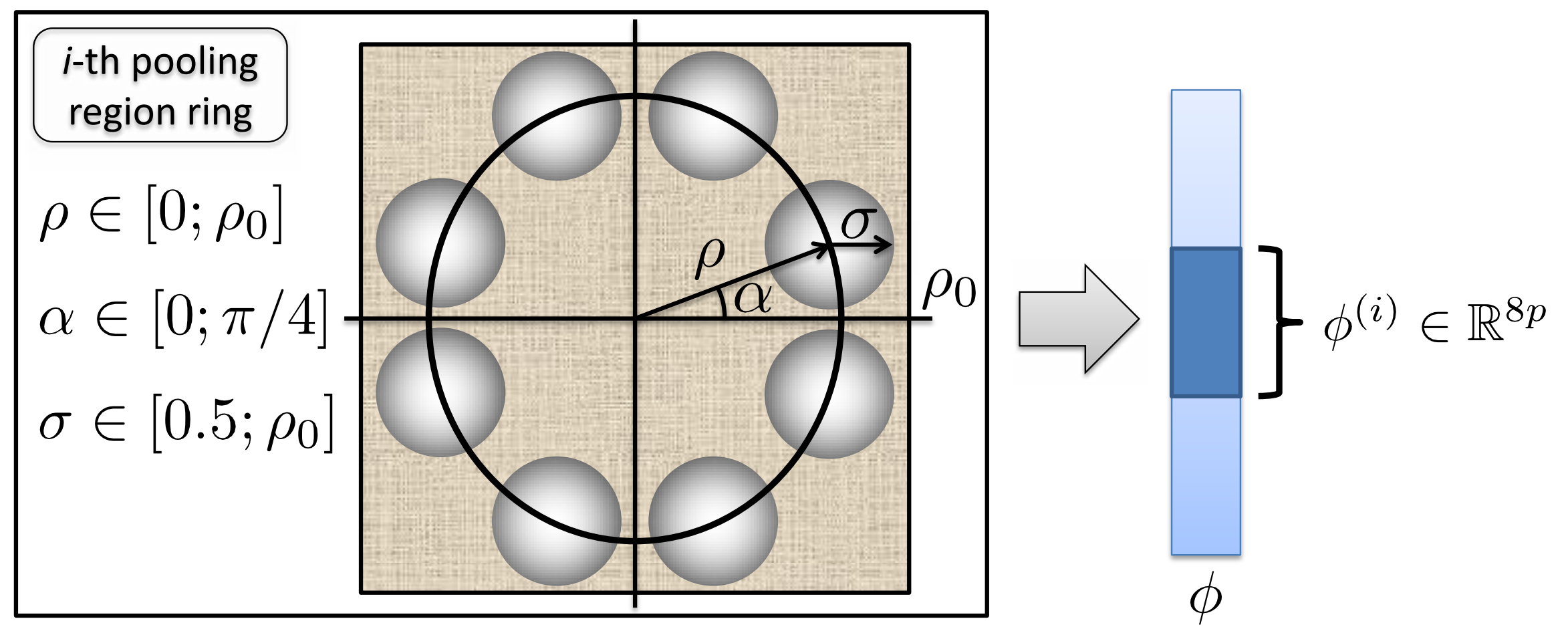
learning W *(panel 4)*

Dimensionality reduction $W\phi$ — Linear projection onto a low-dimensional subspace

Descriptor vector — Can be used directly or quantised to visual words

3. LEARNING POOLING REGIONS

- Candidate Pooling Regions (PRs) are generated by dense sampling of their location and size
- Symmetric configuration: PRs grouped into rings



Gaussian pooling regions, grouped into a ring, are applied to feature channels to produce a part of the descriptor vector $\phi^{(i)}$

- PR learning – selection of a few (≤ 10) PR rings from a large pool (4200 rings)
  - each ring is assigned a non-negative scalar weight w_i
  - squared L^2 distance between descriptors is linear in w
  - sparse weight vector w is learnt

Learning constraints: squared L^2 distance between descriptors of matching keypoint pairs should be smaller than that of non-matching pairs

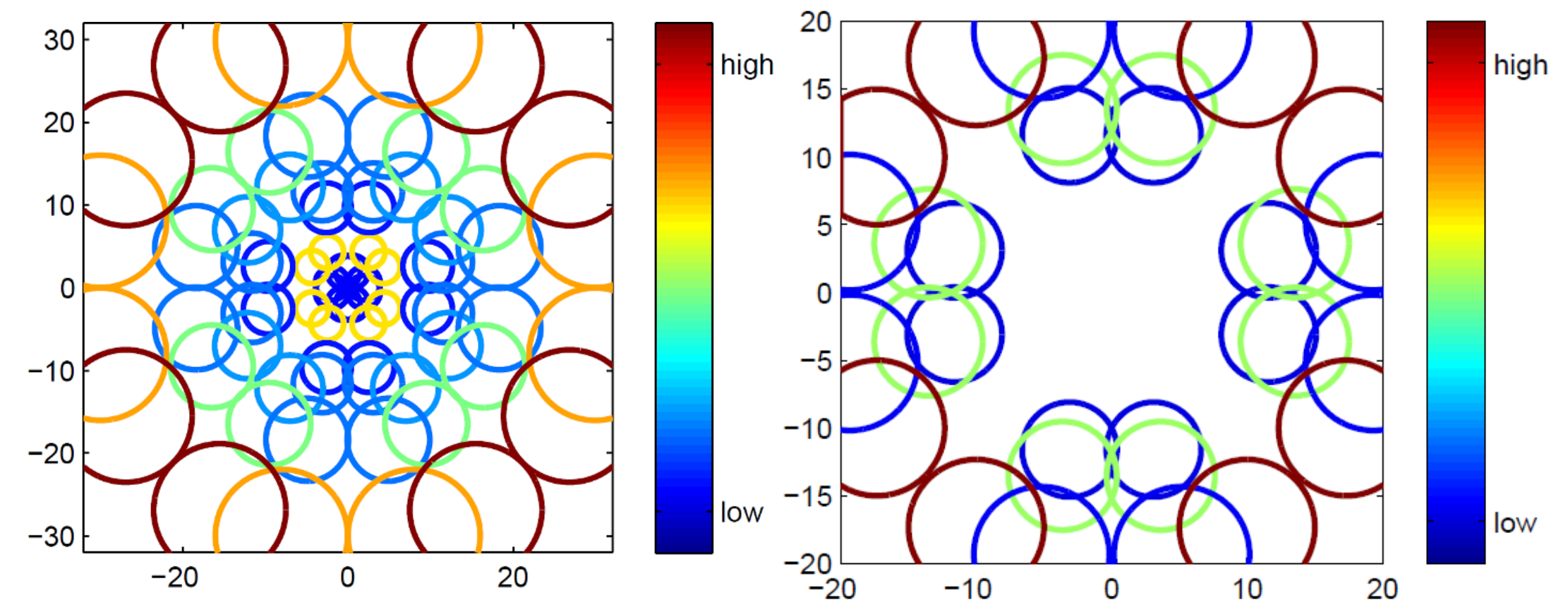
Convex optimisation problem (solved by RDA):

$$\underset{w\ge 0}{\operatorname{argmin}} \underbrace{\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{P},(\mathbf{u},\mathbf{v})\in\mathcal{N}}}_{} \max\left\{w^T\left(\psi(\mathbf{x},\mathbf{y})-\psi(\mathbf{u},\mathbf{v})\right)+1,0\right\}+\mu_1\|w\|_1$$

matching keypoint pairs

non-matching keypoint pairs

$\psi_i(\mathbf{x},\mathbf{y}) = \|\phi^{(i)}(\mathbf{x}) - \phi^{(i)}(\mathbf{y})\|_2^2$



Examples of learnt pooling region configurations (left: 576-D, right: 256-D).

4. LEARNING DIMENSIONALITY REDUCTION

- Linear projection $W \in \mathbb{R}^{m\times n}, m<n$ into lower-dimensional space learnt from the constraints above
- Optimisation over W is not convex, so $A = W^TW$ is optimised instead
- Low-rank projection is enforced by the nuclear norm regularisation $\|A\|_*$ – the sum of singular values
- Nuclear (trace) norm – convex surrogate of rank

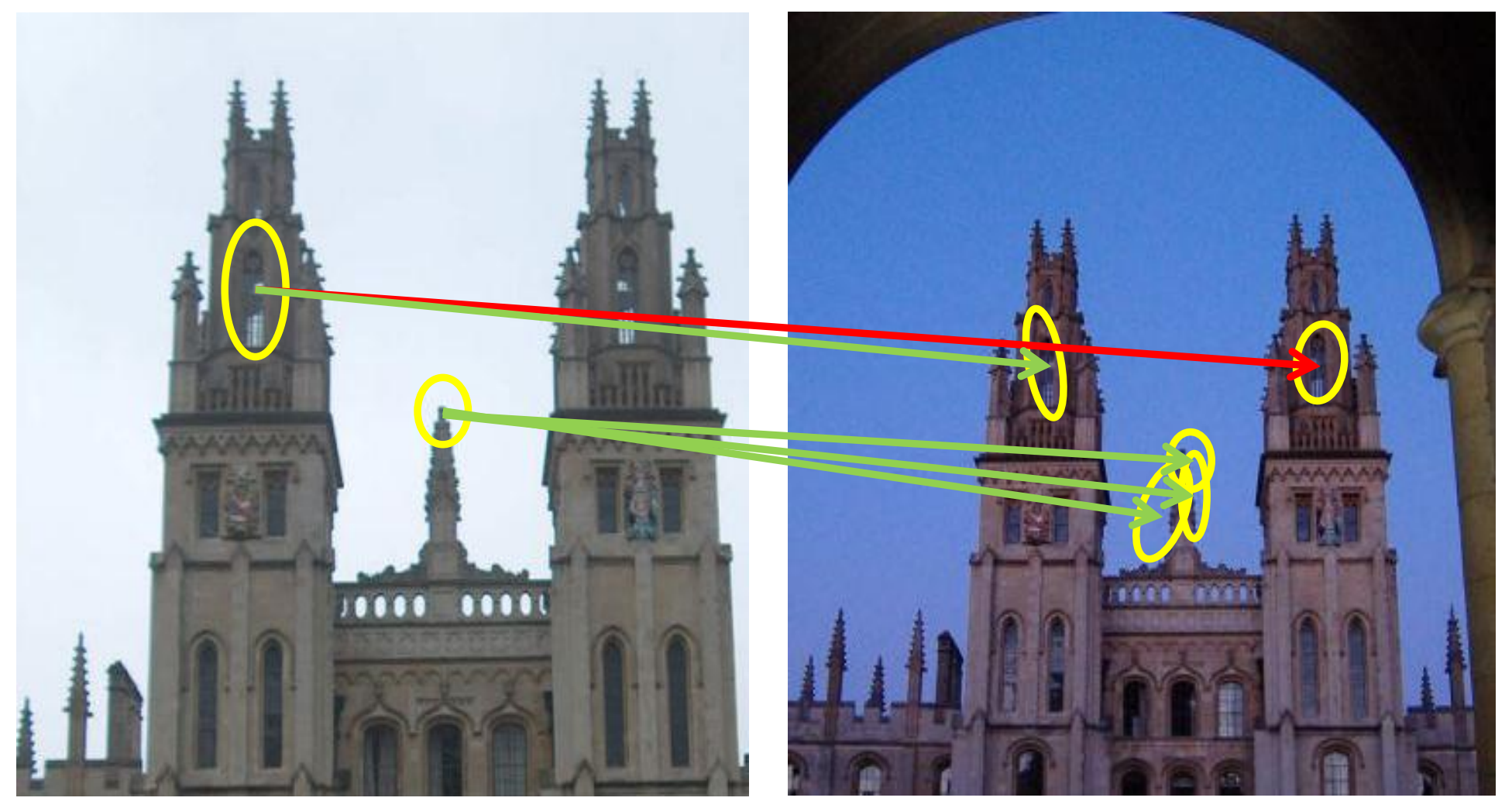
Convex optimisation problem (solved by RDA):

$$\underset{A\succeq 0}{\operatorname{argmin}} \sum_{\substack{(\mathbf{x},\mathbf{y})\in\mathcal{P}\\(\mathbf{u},\mathbf{v})\in\mathcal{N}}} \max\left\{\theta(\mathbf{x},\mathbf{y})^T A\,\theta(\mathbf{x},\mathbf{y}) - \theta(\mathbf{u},\mathbf{v})^T A\,\theta(\mathbf{u},\mathbf{v}) + 1, 0\right\} + \mu_*\|A\|_*$$

difference of descriptor vectors $\theta(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y})$

5. LEARNING FROM WEAK SUPERVISION

- Learning from image datasets with extremely weak annotation: "some (unknown) pairs of images contain a common part" (e.g. Oxford5K)
- Automatic homography estimation using RANSAC
- For each keypoint, a set of putative matches is computed using the affine region overlap criterion



Putative matches (green arrows) are computed from geometry cues. Only the putative match, closest in the current descriptor space, will be used for learning at the next iteration. If confusing non-matches are present, e.g. due to repetitive structure (red arrow), then the keypoint is not used in learning.

- Some keypoints can not be matched based on appearance (due to occlusions, repetitive structure) – modelling matching feasibility with latent variables

Learning constraints: the nearest neighbour of a keypoint, matchable in the descriptor space, should belong to the set of putative matches

Optimisation problem (solved by alternation & RDA):

$$\underset{\eta,b}{\arg\min}\sum_{\mathbf{x}} b(\mathbf{x}) \max\left\{\min_{\mathbf{y}\in P(\mathbf{x})} d_\eta(\mathbf{x},\mathbf{y}) - \min_{\mathbf{u}\in N(\mathbf{x})} d_\eta(\mathbf{x},\mathbf{u}) + 1, 0\right\} + R(\eta)$$

$$\text{s.t. } b(\mathbf{x}) \in \{0,1\},\ \sum_{\mathbf{x}} b(\mathbf{x}) = K$$

matching feasibility indicator

number of matchable training pairs (optimised on the validation set)

pooling region selection or dimensionality reduction model

6. REGULARISED DUAL AVERAGING (RDA)

- Stochastic proximal gradient method well suited for non-smooth objectives with sparsity-enforcing regularisation (e.g. L^1 or nuclear norms)
- Objective: $\min\limits_{w} \frac{1}{T}\sum_{t=1}^{T} f(w, z_t) + R(w)$
- Update: $w_{t+1} = \underset{w}{\operatorname{argmin}}\left(\langle \bar g_t, w\rangle + R(w) + \frac{\beta_t}{t}h(w)\right)$

sub-gradient averaged across iterations

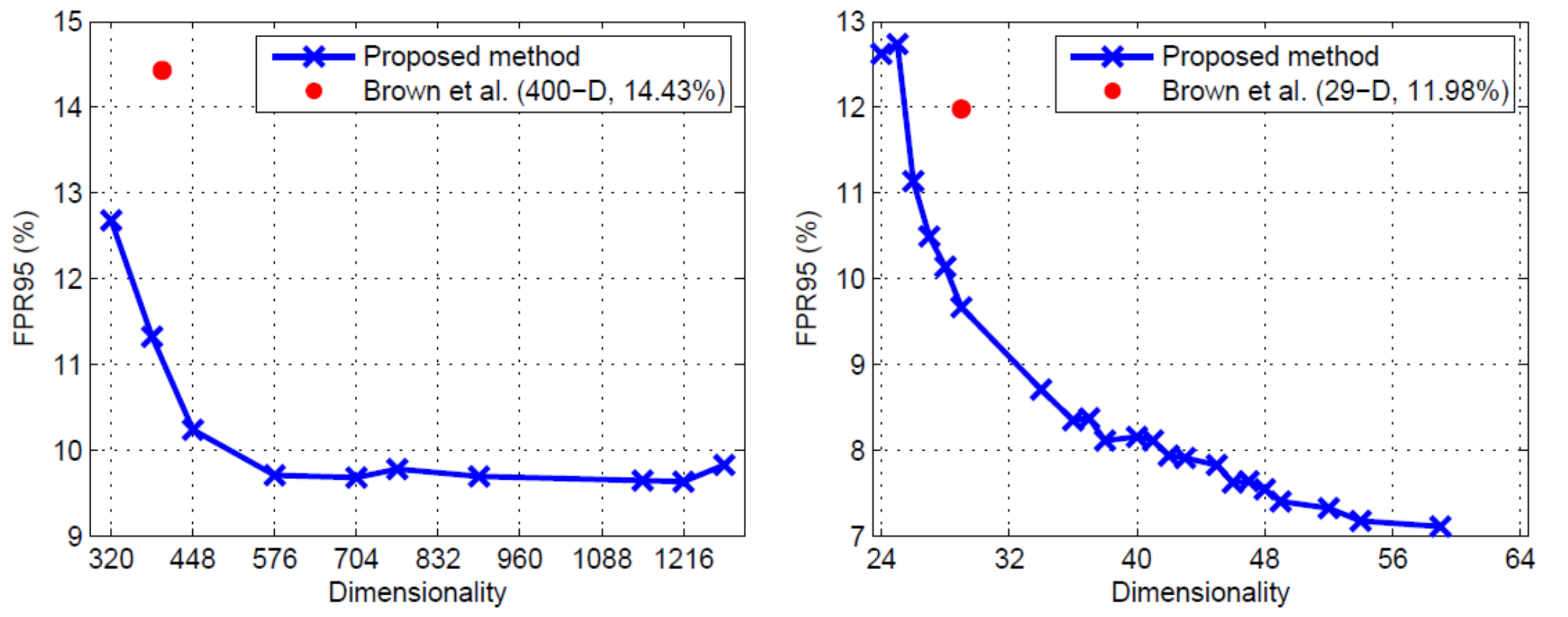
strongly convex proximal function

7. RESULTS: PATCH MATCHING

- Local patches dataset of Brown et al. [PAMI, 2011]
- Measure: false positive rate at 95% recall (FPR95, %)
- **State-of-the-art** performance:

Train set	Test set	Learnt proj., ≤64-D	Learnt proj., low-dim.	Brown *et al.* [2]
Yosemite	Notre Dame	**7.11 (59-D)**	9.67 (29-D)	11.98 (29-D)
Yosemite	Liberty	**16.27 (59-D)**	17.44 (29-D)	18.27 (29-D)
Notre Dame	Yosemite	**10.36 (61-D)**	12.54 (36-D)	13.55 (36-D)
Notre Dame	Liberty	**13.63 (61-D)**	14.51 (36-D)	16.85 (36-D)

Error rate for the learnt descriptors and the method of Brown et al.



Dimensionality vs error rate.
Left: learning pooling regions; right: learning dimensionality reduction.

8. RESULTS: IMAGE RETRIEVAL

- Oxford Buildings and Paris Buildings datasets
- Measure: mean Average Precision (mAP)
- Training on Oxford5K from weak supervision, testing on Oxford5K and Paris6K
- **Outperforms** descriptor learning of Philbin et al. [ECCV, 10] :

Descriptor	mAP			mAP improvement (%)		
	raw	tf-idf	tf-idf+sp.	raw	tf-idf	tf-idf+sp.
Oxford5K						
SIFT	0.784	0.636	0.667	-	-	-
RootSIFT	0.798	0.659	0.703	1.8	3.6	5.4
SIFT + Learnt proj., 120-D	0.802	0.673	0.706	2.3	5.8	5.8
Learnt PR, 256-D	0.819	0.664	0.702	4.5	4.4	5.2
Learnt PR + proj., 115-D	**0.841**	**0.709**	**0.749**	**7.3**	**11.5**	**12.3**
Philbin *et al.* [10], linear	N/A	0.636	0.665	N/A	3.8	2.8
Philbin *et al.* [10], non-linear	N/A	0.662	0.707	N/A	8	9.3
Paris6K						
SIFT	0.691	0.656	0.668	-	-	-
RootSIFT	0.706	0.701	0.710	2.2	6.9	6.3
Learnt PR + proj., 115-D	**0.732**	**0.711**	**0.722**	**5.9**	**8.4**	**8.1**
Philbin *et al.* [10], non-linear	N/A	0.678	0.689	N/A	3.5	3

mAP for learnt descriptors, SIFT, and RootSIFT.

ACKNOWLEDGEMENTS

This work was supported by Microsoft Research PhD Scholarship Program and ERC grant VisRec no. 228180. A. Vedaldi was supported by the Violette and Samuel Glasstone Fellowship.

SUMMARY

Descriptors can be learnt using convex large-margin formulations, leading to state-of-the-art performance

- Pooling region selection using Rank-SVM with L^1 regularisation
- Discriminative dimensionality reduction using large-margin metric learning with nuclear norm regularisation
- Learning under very weak supervision by modelling matching uncertainty with latent variables