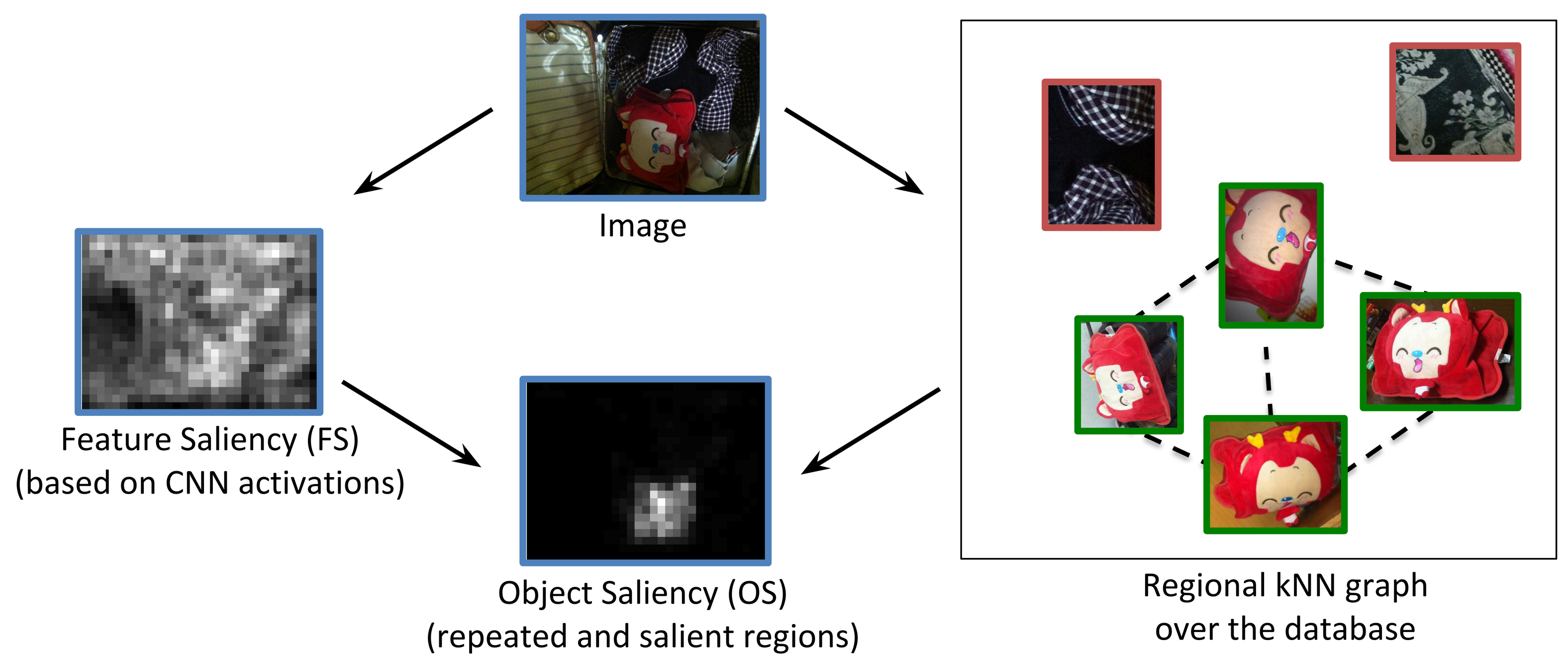
Unsupervised object discovery for instance recognition

Oriane Siméoni[1], Ahmet Iscen[2],
Giorgos Tolias[2], Yannis Avrithis[1], Ondřej Chum[2]
[1]Inria Rennes, [2]Visual Recognition Group, FEE, CTU in Prague

Motivation

- Global CNN descriptors perform well for instance retrieval
- Regional descriptors work better, especially for small objects
  - Higher complexity and memory requirements
- **Solution:** Discover repeating objects, suppress clutter
  - Better global descriptors
- Our method is **fully unsupervised**

Overview

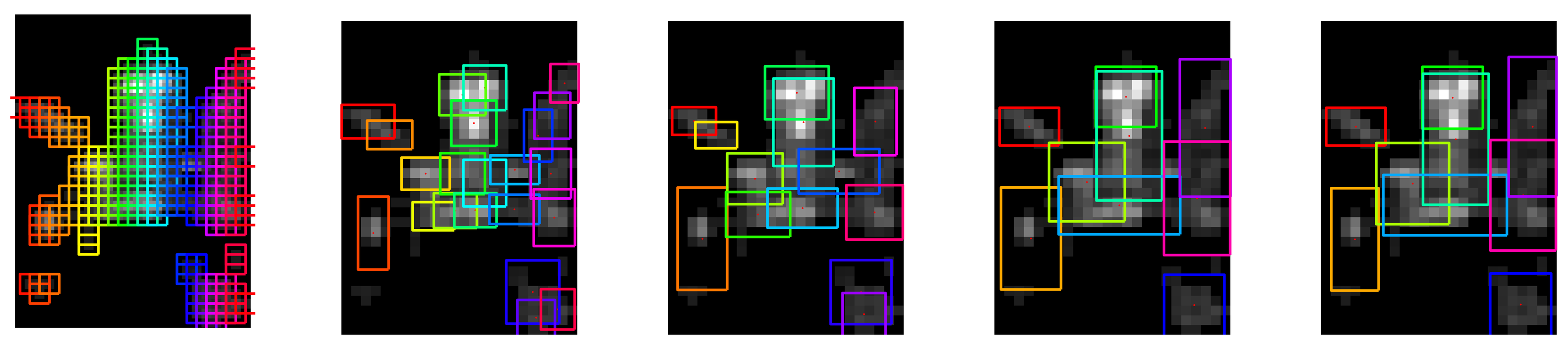


- Detect salient regions on FS and extract regional descriptors
- Construct regional kNN graph over the database
- Regions with many similar neighbors likely contain repeating objects
- Detect regions on OS and aggregate them into a global descriptor

Feature Saliency (FS)

- Create a 2D saliency map of an image from its CNN activations
- Saliency is the normalized sum of the weighted channels
- Weight per channel obtained following CRoW [3]

Region detection: Expanding Gaussian mixture (EGM)



- Detecting a small number of regions from each saliency map
- EGM [1] based on the *expectation-maximization* optimization process
- Dynamically estimates the number of components
- A component is defined as rectangular region in the 2D plane

Database graph construction

- Construct similarity graph between all detected regions from FS maps, with

$$\mathrm{sim}(\mathbf{u},\mathbf{v}) = \begin{cases} (\mathbf{u}^\top\mathbf{v})^\gamma & \text{if } \mathbf{u},\mathbf{v} \text{ are mutual } k\text{-nearest neigbors} \\ 0 & \text{otherwise} \end{cases}$$

- Adjacency matrix W : sparse, symmetric non-negative matrix containing pairwise similarities $w_{ij} = \mathrm{sim}(\mathbf{u}_i, \mathbf{u}_j)$ and zero diagonal
- Symmetrically normalized adjacency matrix:

$$\mathcal{W} := D^{-1/2} W D^{-1/2}$$

where D is the row-wise sum of W

- Regularized graph Laplacian, given $\alpha \in (0,1)$

$$\mathcal{L}_\alpha := (I - \alpha\mathcal{W})/(1-\alpha)$$

Graph centrality [4]

- Centrality g represents the significance of each vertex (region) in the graph
- It is the solution of the linear system:

$$\mathcal{L}_\alpha \mathbf{g} = \mathbf{1}$$

- The solution is obtained by conjugate gradients as in [2]

Object Saliency (OS)

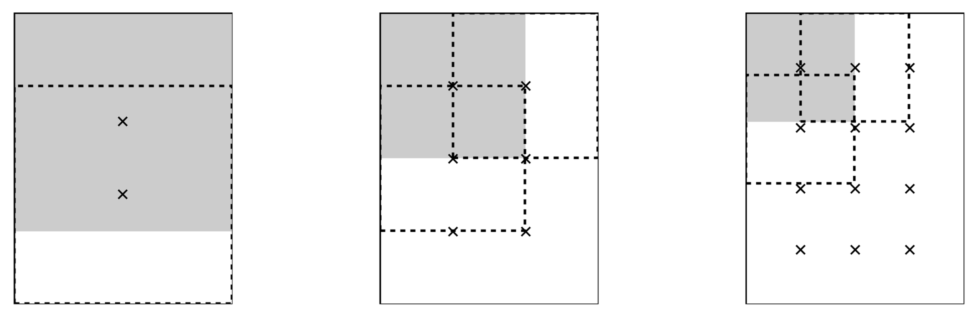
- Object saliency map: reflects relevance to frequent database objects
- Sliding window over the activation map of each image.
- Consider a square patch at each position p and compute saliency S_p:

$$S_p = \overbrace{\hat{F}_p^{\theta}}^{\text{FS at position } p} \sum_{\underbrace{i\in N_p}_{\text{indices of } k\text{-nearest neighbors}}} \overbrace{s(\mathbf{v}_i|\mathbf{x}_p)}^{\text{neighbor region descriptor}} \overbrace{f_i^{\theta}}^{\text{neighbor FS}} \underbrace{g_i^*}_{\text{neighbor centrality score}}$$

- EGM detection from OS maps

Experimental setup: Baselines

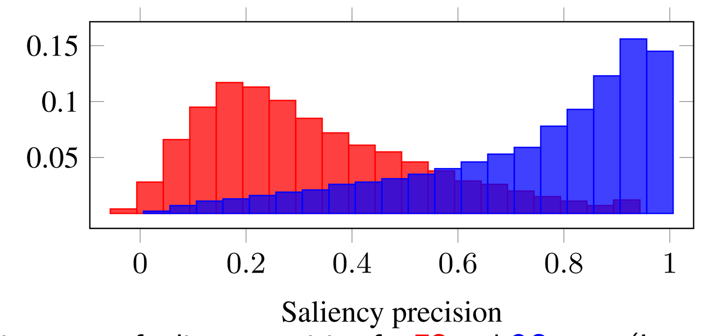
- Uniform: regions selected on a uniform grid at 3 scales.
- Uniform†: same but weighted pooling over activations as in CroW [3].



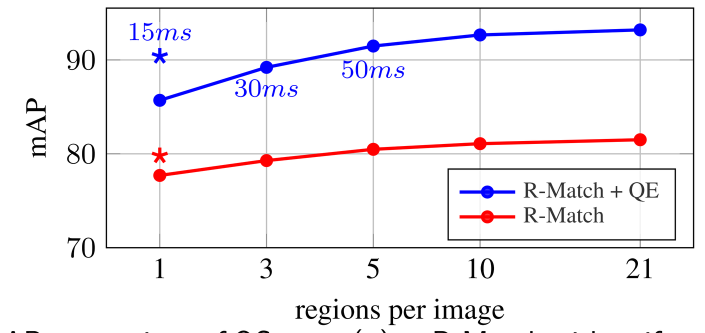
L1 sampling L2 sampling L3 sampling

Results

Method	QE	Instre	Oxford	Oxford105k
MAC	-	48.5	79.7	73.9
Uniform [5]	-	47.7	77.7	70.1
FS.EGM *	-	48.4	77.5	70.2
OS.EGM *	-	50.1	79.6	71.8
OS.EGM-$\triangle^\star$	-	53.7	79.8	71.4
MAC	✓	71.8	87.4	86.0
Uniform [5]	✓	70.3	85.7	82.7
FS.EGM *	✓	71.2	89.8	87.9
OS.EGM *	✓	72.7	**90.4**	**88.0**
OS.EGM-$\triangle^\star$	✓	**75.4**	90.1	84.3



Histogram of saliency precision for FS and OS maps (Instre).

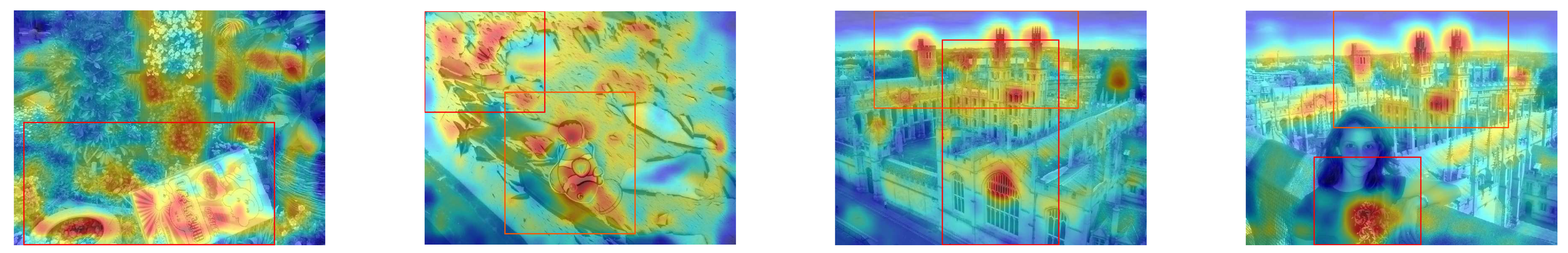


mAP comparison of OS.EGM ($\star$) to R-Match with uniformly sampled regional descriptors (Oxford5k). Text labels refer to query time.

Examples

Image:

FS:



OS:

More OS examples:

References:

[1] Y. Avrithis and Y. Kalantidis. Approximate gaussian mixtures for large scale vocabularies. In *ECCV*, pages 15–28. Springer, 2012.

[2] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum. Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In *CVPR*, 2017.

[3] Y. Kalantidis, C. Mellina, and S. Osindero. Cross-dimensional weighting for aggregated deep convolutional features. *arXiv preprint arXiv:1512.04065*, 2015.

[4] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[5] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *ICLR*, 2016.

Contact: oriane.simeoni@inria.fr