Local Features and Visual Words Emerge in Activations

Oriane Siméoni[1], Yannis Avrithis[1], Ondřej Chum[2]

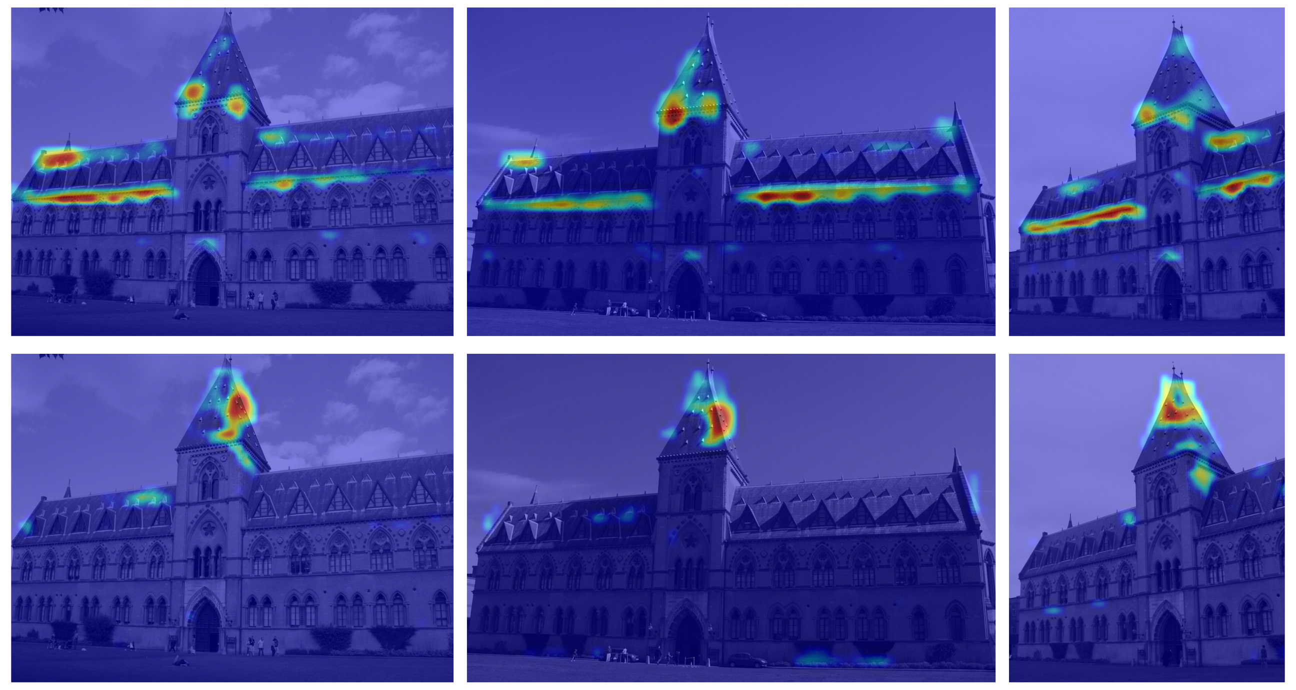
[1]Inria, Univ Rennes,CNRS, IRISA [2]VRG, FEE, Czech Technical University in Prague

Motivation

- **Problem:** spatial verification in large-scale image retrieval
- Global descriptors lose the spatial information in activation maps [1, 6]
- Local descriptors are expensive to store [4]
- **Solution:** detect features directly on activation maps, match them independently per channel
- **No network modification or retraining**
- No local feature detection directly on the input image, no local descriptors, no codebooks

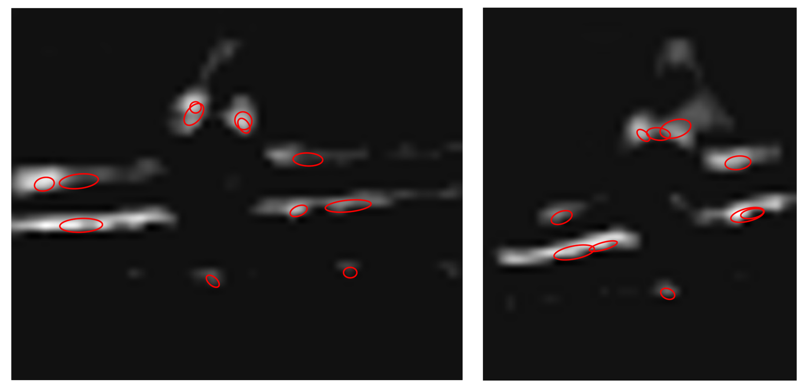
Activation maps

- Activation maps are sparse
- Responses on each channel are corresponding between different views



Local features

Detect local features using MSER [3]



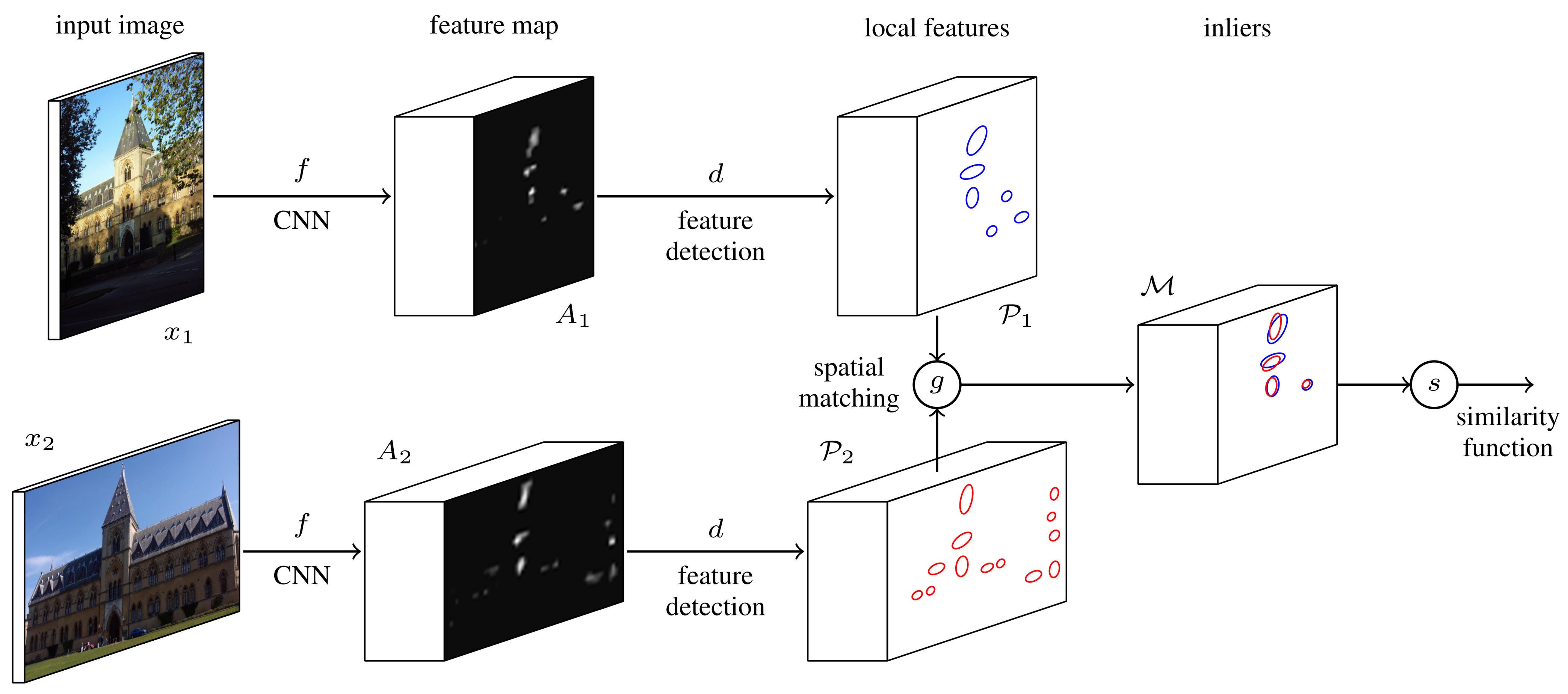
Represent image by a set of features, each specified by
- a scalar strength, pooled over the MSER region (1d)
- mean and covariance matrix of ellipse fitted to the region (5d)
- id of the channel in which the feature was detected (1d)

Spatial matching

- Tentative correspondences only among features in the same channel
- Match tentative correspondences using RANSAC
- **Channel ids play the role of visual words**

Deep Spatial Matching (DSM)

- Get activation maps from the last convolutionnal layer of a CNN
- Approximate tensors by a collection of local features
- Robustly match those features to approximate optimal alignment of tensors



Examples

Diffusion [2] on DSM verified images

- Diffusion based on the nearest neighbor graph of global descriptors
- Ranks images according to manifold similarity
- DSM verified images are a good starting point for diffusion

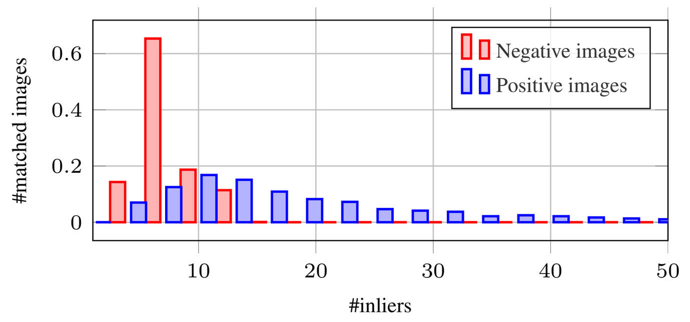
Implementation details

- Images processed at **multiple scales**: 1, $1/\sqrt{2}$, 1/2
- Non-maxima supression by feature strength across channels
- Initial ranking using global descritors, top 100 images re-ranked by DSM

Results: Revisited Oxford and Paris [5]

Method	Medium			
	$\mathcal{R}$Oxf		$\mathcal{R}$Par	
	mAP	mP@10	mAP	mP@10
V	44.8	63.3	65.7	95.0
V+DSM	51.1	77.3	66.2	96.9
R↑	44.4	64.2	69.0	96.4
R↑+DSM	49.6	74.0	69.7	98.4
V+D	48.4	65.2	81.4	95.6
V+DSM+D	61.6	81.0	82.8	97.6
R↑+D	53.8	69.0	85.6	96.3
R↑+DSM+D	60.2	78.9	86.3	96.9

Off-the shelf VGG (V) and ResNet (R). ↑: upsampling; D: diffusion [2]. All results with GeM pooling and supervised whitening.



Distribution of number of **inliers** for positive and negative database images over all queries of $\mathcal{R}$Oxf, using VGG-MAC.

Method	Medium							
	$\mathcal{R}$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
"DELF-HQE+SP" [5]	73.4	88.2	60.6	79.7	84.0	98.3	65.2	96.1
"DELF-ASMK*+SP"→D† [5]	75.0	87.9	68.7	83.6	**90.5**	98.0	**86.6**	98.1
V-MAC*+D	67.7	86.1	56.8	78.6	85.6	97.6	78.6	96.4
V-MAC*+DSM+D	72.0	90.6	59.2	80.1	86.4	98.9	79.3	97.1
R-MAC*↑+D	73.9	87.9	61.3	80.6	89.9	96.1	83.0	95.1
R-MAC*↑+DSM+D	**76.9**	90.7	65.7	83.9	90.1	96.4	84.0	95.3
V-GeM[6]+D	69.6	84.7	60.4	79.4	85.6	97.1	80.7	97.1
V-GeM[6]+DSM+D	72.8	89.0	63.2	83.7	85.7	96.1	80.1	95.7
R-GeM[6]+D	69.8	84.0	61.5	77.1	88.9	96.9	84.9	95.9
R-GeM[6]↑+D	70.1	84.3	67.5	79.0	89.1	97.3	85.0	96.6
R-GeM[6]↑+DSM+D	75.0	89.6	**70.2**	84.5	89.3	97.1	84.8	95.3

State-of-the-art VGG (V) and ResNet (R). ↑: upsampling; *: our re-training; D: diffusion [2]. Results citing [5] as reported in that work, combining DELF [4], ASMK* and HQE. SP: spatial matching; D†: diffusion on graph obtained by [1].

References

[1] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *IJCV*, 124(2):237–254, Sep 2017.

[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] J. Matas, O. Chum, U. Martin, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[4] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017.

[5] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018.

[6] F. Radenović, G. Tolias, and O. Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. PAMI*, 2018.

Contact: oriane.simeoni@inria.fr