Indoor Scene Recognition using Task and Saliency-driven Feature Pooling

Marco Fornoni[1][2]
http://www.idiap.ch/~mfornoni

Barbara Caputo[2]
http://www.idiap.ch/~bcaputo

[1] Ecole Polytechnique Fédérale, Lausanne (EPFL)
Lausanne, CH

[2] Idiap Research Institute
Martigny, CH

Indoor scene recognition is as of today one of the most challenging open problems in visual place categorization. Since the seminal works of Oliva and Torralba [5] and Lazebnik et al. [3], the mainstream approach to scene recognition has been based on global, appearance-based image representations, enriched with spatial information. This approach, in various forms, has given good results for the outdoor place recognition problem, but proved to be inadequate when dealing with indoor scenes [6]. In indoor environments, indeed, the location of meaningful regions and objects varies drastically within each category. Also, the close-up distance between the camera and the subject makes the variations due to viewpoint changes even more severe. In this scenario, it becomes crucial how low-level features are spatially pooled to get the final image description, especially for the robustness of the representation. In this work we investigate this issue and propose to combine a simple spatial encoding, with a saliency-driven perceptual pooling designed to capture structural properties of the scenes, independently from their position in the image.

Saliency-driven perceptual pooling The traditional spatial encodings are designed to capture the spatial regularities in the scenes. We would instead like to let visual-structures emerge from the data, regardless of their exact position in the imaged scenes. Specifically, we are aiming to obtain a segmentation (R_1, R_2) such that R_2 captures the area of the image with a richer informative content (i.e., a high number of visual word responses), leaving to R_1 the task to collect the statistics of the remaining part. This is obtained by first computing a saliency map for each image, and subsequently using the median saliency value $\bar{s}$ of the image, to segment it in two regions: the most and least salient 50%.
To compute the saliency map, we test two approaches:

- **Itti**. The classic approach described in [2]
- **SIFT Saliency**. A novel saliency operator, directly and solely using the precomputed SIFT features to estimate the saliency map

For the latter saliency function (SIFT), we build on the work of Bruce and Tsotsos [1]. In their proposal, the probability of each pixel is locally estimated by non-parametrically fitting a distribution over the ICA projection of the RGB values of the image. Similar to [1], after computing the ICA projection $\bar{\mathbf{X}} = [\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_N]^T$ of the SIFT description $\mathbf{X}$ of an image, we estimate the probability of the j-th dimension of a descriptor i as:

$$p(\bar{\mathbf{x}}_{i,j}) = \frac{1}{N} \sum_{k=1}^{N} K(\bar{\mathbf{x}}_{i,j} - \bar{\mathbf{x}}_{k,j}), \tag{1}$$

where $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$ is a one-dimensional standard Gaussian kernel. The saliency of the local descriptor $\bar{\mathbf{x}}_i$ is then computed as:

$$s(\bar{\mathbf{x}}_i) = -\sum_{j=1}^{D} \log \bar{\mathbf{x}}_{i,j} \tag{2}$$

and the final saliency map is obtained by computing the responses for all the SIFT descriptors of the image, followed by a smoothing operation.

Task-driven spatial pooling Indoor scenes are designed to support human actions and humans have a limited range of spatial mobility. For example, humans cannot easily move from the floor to the ceiling, or access facilities if they are disposed too low, or too high in the room. This reduces the spatial variability of indoor scenes to lie mostly on the horizontal axis. Given this prior, we expect that by pooling features in horizontal bands we will be able to capture the most consistent spatial patterns in indoor scenes. We instead expect less robust results by pooling descriptors in vertical bands.
To verify this intuition we thus compare the following pooling schemes:

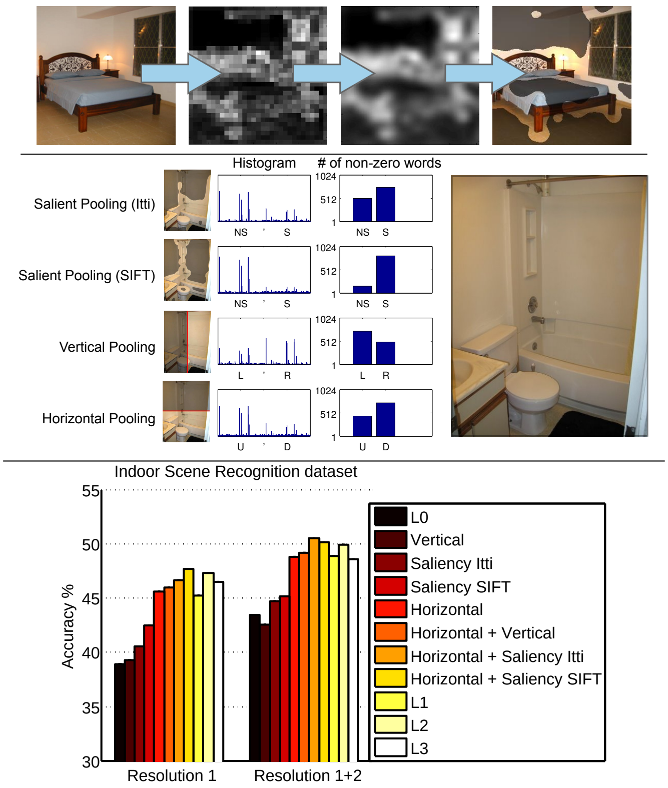

Figure 1: Top: Computation of a SIFT saliency map and resulting segmentation. Middle: Histograms obtained with different pooling techniques and number of non-zero visual words in each of the two halves of the histograms: non-salient (NS) and salient (S), left (L) and right (R), up (U) and down (D). Bottom: Performances of the different pooling strategies on the Indoor Scene Recognition [6] dataset.

- Horizontal-bands pooling. In this settings R_1 consists of the upper 50% of the image, while R_2 is its complement
- Vertical-bands pooling. In this case R_1 consists of the left-side 50% of the descriptors, and R_2 is again its complement

We performed experiments on three widely used scene recognition datasets: the Indoor Scene Recognition (*ISR*) [6], the 15-Scenes [3] and the 8-Sports [4] datasets. A visualization of the pooling strategies, together with the resulting histograms and a performance evaluation on the ISR dataset are shown in Fig. 1. We see that the salient pooling strategies perform better than the vertical one (+8.1% relative improvement). Moreover, when combined with the horizontal pooling, they always outperform the Horizontal + Vertical and the $L1$ spatial pyramid baselines, being also competitive with much higher dimensional representations, like $L2$ / $L3$.

[1] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, 2006.

[2] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 2002.

[3] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[4] L.J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[6] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.