Beyond Spatial Pooling: Fine-Grained Representation Learning in Multiple Domains

Chi Li, Austin Reiter, Gregory D. Hager
Department of Computer Science, Johns Hopkins University.

The core challenge of object recognition is to create representations that are robust to appearance variations. Recent advances in convolutional architectures [1, 2, 4, 5] have achieved success in learning object representations with minor scale and shift invariances. *Spatial Pooling*, which groups local features within spatial neighborhoods, is a key element to achieve those invariance properties. However, creating object representations that are robust to changes in viewpoint while capturing local visual details continues to be a challenge. In this work, we formulate a probabilistic framework for analyzing the performance of pooling in the context of convolutional architectures and propose a simple but effective solution of pooling beyond spatial domain using adaptive scales of filters, to address the feature misalignment problem. Our major contributions are three-fold as follows:

1. A probabilistic framework mathematically explains how the pooling granularity, filter scale and pooling domain affects the pooled features in terms of the overall discrimination and invariance.

2. A novel multi-scale and multi-domain pooling algorithm is described to exploit adaptive filter scales and invariant pooling domains for fine-grained representation learning.

3. A new 'JHUIT-50' dataset including 50 industrial objects and hand tools is presented with a novel experiment setting.

One pooling theory was proposed by Boureau [3] in the context of hard-assignment coding and i.i.d Bernoulli distributions within pooling regions. These conditions restrict its generalization.

Consider a pooling domain $S = \{s_1, \cdots, s_N\}$ where pooling state s_j with $1 \leq j \leq N$ is a coordinate over which pooling takes place. For example, in the case of RGB-D data, S can be a set of spatial coordinates or color values, corresponding to spatial and color domains. Considering a random sampling of poses under 3D transformation $\mathcal{T}$ for object o_p, let $X^p = (x^p_{11}, ..., x^p_{jk}, ..., x^p_{NK})$ denotes the random vector of filter responses where each $x^p_{jk} = (s_j, d_k)|o_p$ describes the distribution of the activation strength of the kth filter d_k at s_j given o_p.

We measure the variability of X^p with an invariance score J^p:

$$J^p = E(||X^p - \widetilde{X}^p||_2^2) = \sum_{j=1}^{N}\sum_{k=1}^{K} 2\mathrm{Var}(x^p_{jk}) \tag{1}$$

where $\widetilde{X}^p$ is an alias for X^p with $P(X^p) = P(\widetilde{X}^p)$.

The above probabilistic formulation yields three major conclusions:

1. Pooling Granularity: As pooling granularity changes from fine to coarse levels, pooled features have better invariance but less discrimination.

Let $\mathcal{R} = \{R_1, \cdots, R_M\}$ be a partition of S (overlapping case is also analyzed in the paper) and assume max pooling is used (the results are also applied to average pooling). Define a new random variable $y_{ik} = \max_{s_j \in R_i} x_{jk}$ for pooled filter response in pooling region R_i. Analogous to X^p, we then define the random vector $Y^p_{\mathcal{R}} = (y^p_{11}, y^p_{12}, \cdots, y^p_{MK})$. $J^p_{\mathcal{R}}$ is the invariance score of the pooled representation $Y^p_{\mathcal{R}}$. We can then prove the following result using the fact that $\mathrm{Var}(\max_i X_i) \leq \sum_i \mathrm{Var}(X_i)$:

$$J^p_{\mathcal{R}} = \sum_{k=1}^{K}\sum_{i=1}^{M} 2\mathrm{Var}(y^p_{ik}) \leq \sum_{k=1}^{K}\sum_{j=1}^{N} 2\mathrm{Var}(x^p_{jk}) = J^p \tag{2}$$

On the other hand, the discrimination term $(\|\Delta E_{\mathcal{R}}\|_2^2)$ tends to decrease with high probability.

2. Adaptive Filter Scales: Small-scale filters achieve better invariance than the large-scale ones in fine-grained pooling.

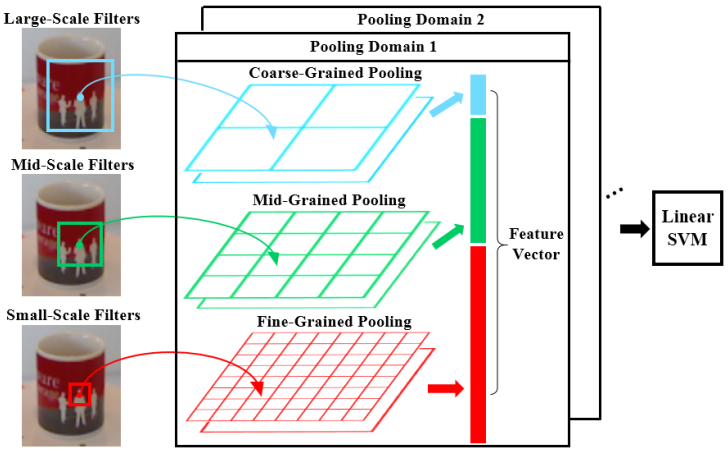

Figure 1: Overview of multi-scale and multi-domain pooling architecture.

The distribution of filter response can be decomposed as follows:

$$P(x^p_{jk}) = P(d_k|s_j, o_p)P(s_j|o_p) \tag{3}$$

This implies that $\mathrm{Var}(x^p_{jk})$ is positively proportional to $\mathrm{Var}(d_k|s_j, o_p)$. Reducing $\mathrm{Var}(d_k|s_j, o_p)$ can be interpreted as choosing filters that have smaller variance across the pooling domain S. Therefore small-scale filters are preferred than large-scale ones because the value changes of local regions are less than large areas in convolution. However, large-scale filters are prone to create better discrimination, which is favored in coarse-grained pooling.

3. Invariant Pooling Domain: Pooling domains insensitive to transformations obtain better invariance in fine-grained pooling.

Alternatively, $P(x^p_{jk})$ can be decomposed as follows:

$$P(x^p_{jk}) = P(s_j|d_k, o_p)P(d_k|o_p) \tag{4}$$

Similarly, we could decrease $\mathrm{Var}(x^p_{jk})$ by reducing $\mathrm{Var}(s_j|d_k, o_p)$, which guides us to construct a pooling domain where appearance features have better alignments at each s_j. Considering 3D transformations, spatial layouts of the transformed object samples change sharply while color configurations are typically aligned across different poses. This fact motivates us to exploit the color domain as an example of an invariant domain in this study.

Algorithm: The three theoretical views above directly lead to the design of the multi-scale and multi-domain pooling architecture illustrated in Fig. 1. Pooled features from fine to coarse pooling levels across different domains are concatenated together to generate the final representation, and a linear SVM is used for the classification.

[1] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *NIPS*, 2011.

[2] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Multipath sparse coding using hierarchical matching pursuit. In *CVPR*. IEEE, 2013.

[3] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *ICML*, 2010.

[4] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *ICML*, 2014.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.