Deep filter banks for texture recognition and segmentation

Mircea Cimpoi[1], Subhransu Maji[2], Andrea Vedaldi[1]
[1]Department of Engineering Science, Oxford University [2]University of Massachusetts, Amherst

Texture is ubiquitous and provides useful cues of material properties of objects and their identity, especially when shape is not useful. Hence, a significant amount of effort in the computer vision community has gone into recognizing texture via tasks such as texture perception and description, material recognition, segmentation, and even synthesis.

Perhaps the most studied task in texture understanding is the one of material recognition, as captured in benchmarks such as CuRET [4], KTH-TIPS [1], and, more recently, FMD [13]. However, while at least the FMD dataset contains images collected from the Internet, vividly dubbed "images in the wild", all these datasets make the simplifying assumption that textures fill images. Thus, they are not necessarily representative of the significantly harder problem of recognizing materials in natural images, where textures appear in clutter. Building on a recent dataset collected by the computer graphics community, the **first contribution** of this paper is a *first large-scale analysis of material and perceptual texture attribute recognition and segmentation in clutter* (Fig. 1).

Motivated by the challenge posed by recognizing texture in clutter, we develop a new texture descriptor. Texture representations based on orderless pooling of local image descriptors set the state-of-the-art in many image understanding tasks, not only for textures, but for objects and scenes too. Currently, however, Convolutional Neural Networks (CNNs) have emerged as the new state-of-the-art for recognition, exemplified by remarkable results in image classification [9], detection [5] and segmentation [7] on a number of widely used benchmarks. Importantly, CNNs pre-trained on such large datasets have been shown [2, 5, 10] to contain general-purpose feature extractors, transferrable to many other domains.

Domain transfer in CNNs is usually achieved by using as features the output of a deep, fully-connected layer of the network. From the perspective of textures, however, this choice has three drawbacks. The first one (I) is that, while the convolutional layers are akin to non-linear filter banks, the fully connected layers capture their spatial layout. While this may be useful for representing the shape of an object, it may not be as useful for representing texture. A second drawback (II) is that the input to the CNN has to be of fixed size to be compatible with the fully connected layers, which requires an expensive resizing of the input image, particularly when features are computed for many different regions [5, 6]. A third and more subtle drawback (III) is that deeper layers may be more domain-specific and therefore potentially less transferrable than shallower layers.

The **second contribution** of this paper is to propose FV-CNN, a *pooling method that overcomes these drawbacks*. The idea is to regard the convolutional layers of a CNN as a filter bank and build an orderless representation using Fisher vector [11] as a pooling mechanism, as is commonly done in the bag-of-words approaches. Although the suggested change is simple, the approach is remarkably flexible and effective. First, pooling is orderless and multi-scale, hence suitable for textures. Second, any image size can be processed by convolutional layers, avoiding costly resizing operations. Third, convolutional filters, pooled by FV-CNN, are shown to transfer more easily than fully-connected ones even without fine-tuning. While others [6, 8] have recently proposed alternative pooling strategies for CNNs, we show that our method is more natural, faster and often significantly more accurate.

The **third contribution** of the paper is a *thorough evaluation of these descriptors* on a variety of benchmarks, from textures to objects. In textures, we evaluate material and describable attributes recognition and segmentation on new datasets derived from OpenSurfaces. When used with linear SVMs, FV-CNN improves the state of the art on texture recognition by a significant margin, obtaining 79.8% accuracy on the Flickr material dataset (previous best 66.7% [3]) and 81.8% accuracy on KTH-2b (previous best of 76.2% [3]). Like textures, scenes are also weakly structured and a bag-of-words representation is effective. FV-CNN obtains 81% accuracy on the MIT indoor scenes dataset [12], significantly outperforming the current state-of-the-art of 70.8% [16]. What is remarkable is that, where [16] finds that CNNs trained on scene recognition data perform better than CNNs trained on an object domain (ImageNet), when used in FV-CNN not only there is an overall performance improvement, but the domain-specific advantage is entirely removed. This indicates that FV-CNN are in fact better at domain transfer. Our method also matches the previous best in PASCAL VOC 2007 classification dataset providing measurable boost over CNNs and is closely following competitor methods on CUB 2010-2011 datasets when no ground-truth object bounding boxes are given: FV-CNN achieves 66.7% accuracy on the CUB 2010-2011 dataset [14] requiring *only* image labels for training and 73.0% accuracy given the object bounding box, closely matching the current best methods that additionally require landmark annotations during training (76.37% [15]).

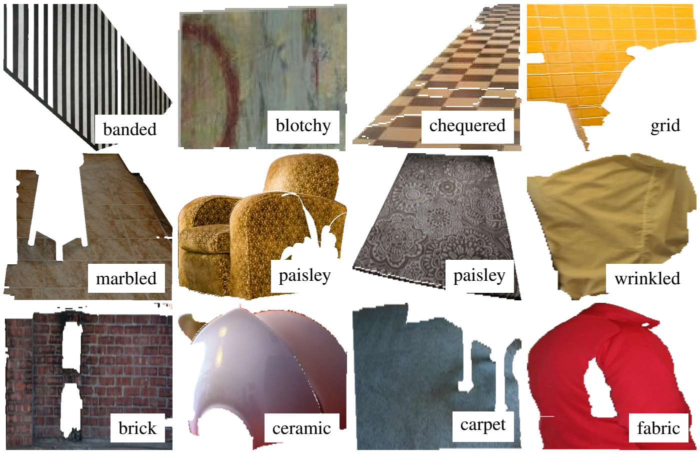

Figure 1: **Texture recognition in clutter**. Example of top retrieved texture segments by attributes (top two rows) and materials (bottom) in the OpenSurfaces dataset.

[1] B. Caputo, E. Hayman, and P. Mallikarjuna. Class-specific material categorisation. In *ICCV*, 2005.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC*, 2014.

[3] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proc. CVPR*, 2014.

[4] K. J. Dana, B. van Ginneken, S. K. Nayar, and J. J. Koenderink. Reflectance and texture of real world surfaces. *ACM Transactions on Graphics*, 18(1):1–34, 1999.

[5] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014.

[6] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *Proc. ECCV*, 2014.

[7] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *Computer Vision–ECCV 2014*, pages 297–312. Springer, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proc. ECCV*, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[10] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *Proc. CVPR*, 2014.

[11] F. Perronnin and C. R. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

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

[13] L. Sharan, R. Rosenholtz, and E. H. Adelson. Material perceprion: What can you see in a brief glance? *Journal of Vision*, 9:784(8), 2009.

[14] P. Welinder, S. Branson, T. Mita, C. Wah, and F. Schroff. Caltech-ucsd birds 200. Technical report, Caltech-UCSD, 2010.

[15] N. Zhang, J. Donahue, R. Girshickr, and T. Darrell. Part-based R-CNNs for fine-grained category detection. In *Proc. ECCV*, 2014.

[16] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Proc. NIPS*, 2014.

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