Mid-level Deep Pattern Mining

Yao Li[1,2], Lingqiao Liu[1], Chunhua Shen[1,3], Anton van den Hengel[1,3]
[1]University of Adelaide. [2]NICTA. [3]Australian Centre for Robotic Vision.

Mid-level visual elements, which are clusters of image patches rich in semantic meaning, were proposed by Singh *et al.* [9] with the aim of replacing low-level visual words. In this pioneering work, mid-level visual elements must meet two requirements, that is, representativeness and discriminativeness. Representativeness means mid-level visual elements should frequently occur in the target category, while discriminativeness implies that they should be visually discriminative against the natural world. The discovery of mid-level visual elements has boosted performance in a variety of vision tasks, including image classification, action recognition, discovering stylistic elements, geometry estimation and 2D-3D alignment.

In this paper, building on the well-known association rule mining, we propose a pattern mining algorithm, *Mid-level Deep Pattern Mining* (MDPM), to study the problem of mid-level visual element discovery. This approach is particularly appealing because the specific properties of activation extracted from the fully-connected layer of a Convolutional Neural Network (CNN) allow them to be seamlessly integrated with association rule mining, which enables the discovery of category-specific patterns from a large number of image patches . Also, we find that two requirements of mid-level visual elements, representativeness and discriminativeness, can be easily fulfilled by association rule mining. When we visualize image patches with the same pattern (mid-level visual element in our scenario), it turns out they are not only visually similar, but also semantically consistent (See Fig. 1).

The main steps the proposed MDPM algorithm are as follows.

1. **Transaction creation.** As we aim to discover patterns from image patches through pattern mining, *an image patch is utilized to create one transaction*. We treat *each dimension index of CNN activation* as an item (4096 items in total). Each transaction is then represented by *the dimension indices of the k largest magnitudes of the corresponding image patch*. Following the work of [7], at the end of each transaction, we add a *pos* (*neg*) item if the corresponding image patch comes from the target category (natural world).
2. **Association rule mining.** Given the transaction database $\mathcal{D}$, we use the Aprior algorithm [1] to discover a set of patterns $\mathcal{P}$ through association rule mining. Each pattern $P \in \mathcal{P}$ discovered by association rule mining must satisfy the following two criteria:

$$\text{supp}(P) > \text{supp}_{\text{min}}, \quad (1)$$

$$\text{conf}(P \rightarrow pos) > \text{conf}_{\text{min}}, \quad (2)$$

where supp_{min} and conf_{min} are thresholds for the support value and confidence.

We apply the proposed MDPM algorithm to the task of image classification, which includes the following steps.

1. **Retrieving mid-level visual elements.** Given a discovered pattern $P \in \mathcal{P}$, the corresponding mid-level visual element is the set of image patches sharing the this pattern, which can be retrieved efficiently through an inverted index.
2. **Image feature generation.** Given the retrieved mid-level visual elements, we then merge some over-lapping visual elements and train detectors simultaneously. Similar to previous works [2, 5, 9], we select an subset of detectors and evaluate them on a new image at multiple scales. For each scale, we take the max score per detector per region encoded in a 2-level spatial pyramid. The final feature vector is the outcome of max pooling on the features from all scales.

We evaluate our whole approach on two benchmark image classification datasets, PASCAL VOC 2007 [3] and MIT Indoor [8].

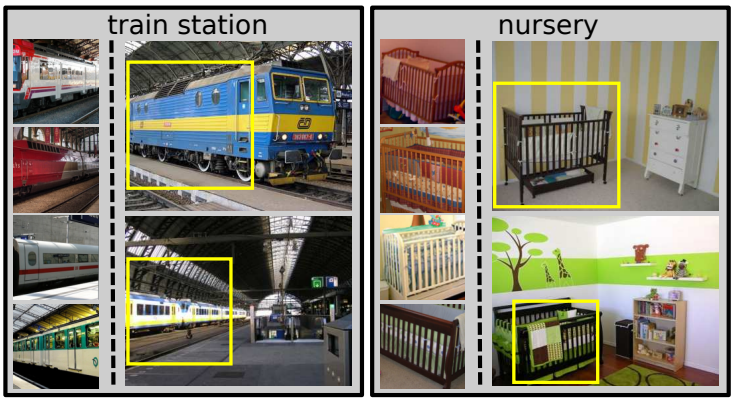

Figure 1: Discovered mid-level visual elements and their corresponding detections on test images on the MIT Indoor dataset.

When using 20 element detectors per category, our approach 68.24% mean accuracy on MIT Indoor dataset. Compared with the work of Doersch *et al.* [2] which achieved best performance among previous mid-level visual elements algorithms, our approach uses an order of magnitude fewer elements than [2] (20 *vs.* 200) while outperforming it by over 4 percent in accuracy. Our best performance, 69.69% mean accuracy, is achieved when the number of element detectors used per category is increased to 50, outperforming recent works using CNN features [4, 6, 10].

On the PASCAL VOC 2007 dataset, our approach achieves 75.2% mean average precision (mAP), significantly outperforming the baseline that using CNN activations as a global feature (67.3%).

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[2] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, pages 494–502, 2013.

[3] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[4] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, pages 392–407, 2014.

[5] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, pages 923–930, 2013.

[6] L. Liu, C. Shen, L. Wang, A. Hengel, and C. Wang. Encoding high dimensional local features by sparse coding based fisher vectors. pages 1143–1151, 2014.

[7] T. Quack, V. Ferrari, B. Leibe, and L. J. V. Gool. Efficient mining of frequent and distinctive feature configurations. In *ICCV*, pages 1–8, 2007.

[8] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420, 2009.

[9] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, pages 73–86, 2012.

[10] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. 2014.

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