HIERARCHICAL MIX-POOLING AND ITS APPLICATIONS TO BIOMEDICAL IMAGE CLASSIFICATION

Siyamalan Manivannan, Ruixuan Wang, Emanuele Trucco

CVIP, School of Science and Engineering, University of Dundee, UK

ABSTRACT

This paper introduces *Hierarchical Mix-pooling* (HMP), a translation-invariant image representation improving the discriminative power of pooling representations by capturing intermediate-size structure information in images. HMP consists of two levels, one traditional pooling (e.g., sum pooling) applied to intermediate-size regions to collect the statistics of local features, and one different pooling (e.g., max pooling) collecting statistics of the previously region-based pooled results. Classification experiments show that HMP considerably improves accuracies with much smaller sizes of dictionaries compared to traditional pooling. The superior performance of HMP is confirmed by experiments with different local features and classifiers on two public biomedical datasets (ICPR HEp-2 cells and IRMA radiology).

1. INTRODUCTION

Sparse coding (SC) [1] has shown its effectiveness as a feature encoding method in various applications in medical image analysis, for instance colonoscopy image classification [2]. Traditional feature encoding based on SC consists of two stages: *coding* and *pooling*. In the coding stage, each local feature (e.g., SIFT, extracted from a small image patch of size 16×16 pixels) is represented by a linear combination of a small set of *codewords* from a dictionary learned in advance (e.g., by clustering). As a result, the local feature can be represented by a sparse vector (the number of non-zero vector components is small) with each vector component corresponding to one codeword. Then in the *pooling* stage, specific statistics of all the sparse vectors are generated to represent the whole image (see Section 2 for detail).

While several studies have tried to improve the robustness of coding, for example, by constraining that similar local features should have similar SC representations [3] , more studies focus on the pooling stage. Traditional pooling techniques (*sum* or *max*; see Section 2) [4] and their variations [5] extract statistical information from all the sparse vectors over the entire image, without considering arguably useful information on the spatial layouts of local features in the image. To solve this issue, *region-based pooling* was proposed, which firstly divides an image into fixed spatial pyramid (SPM) regions, and then the pooled sparse codes from each region are concatenated to get the final image representation [1]. Region-based versions include learning a set of variable size rectangular regions instead of fixed ones [6], applying weights to different regions based on saliency [7], and assigning each local feature to multiple SPM regions with weights [8]. Since region-based pooling encodes location information via direct concatenation of region representations, the final image representation is not invariant to translation. This reduces its applicability when used to represent images which do not have a natural orientation (e.g., ‘up’ in outdoor scenes), as for instance cell images (Figure 1). On the other hand, traditional pooling can avoid encoding the large-scale spatial layout information, but it only captures the statistics of small-size regions (e.g., 16×16 pixels) and not the information about any intermediate-size (e.g., 64×64 pixels) structures in the images. Higher discriminative power for image classification could be achieved by encoding properties of intermediate spatial extent. This paper proposes a simple but effective pooling method, called *Hierarchical Mix-pooling (HMP)*, to achieve a translation-invariant image representation capturing intermediate structure information.

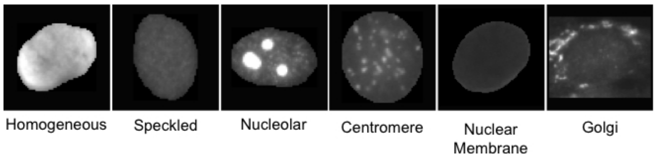


Fig. 1: Example images from the HEp2 cells image dataset.

Recently a pooling approach for deep convolutional neural nets (CNN) called the *mix-pooling* has been proposed in [9]. Mix-pooling tries to reduce the over-fitting problems which is often encountered in CNN by a weighted combination of the sum and max pooling. However, this approach when applied to the traditional dictionary-based approaches also have the same limitations as with the sum and max pooling. Unlike [9], HMP is a two-level approach, therefore it can capture intermediate-size structure information.

Inspired by CNN, various hierarchical SC approaches have been proposed, where coding and pooling are applied in each layer of a multi-level architecture [10, 11]. These hierarchical approaches do capture intermediate structures, but they also increase the complexity of encoding tremendously because a unique dictionary is learned at each layer based on the pooled representation from the previous layer. Unlike hierarchical approaches, the HMP proposed in this paper only needs to learn a single dictionary for two levels of pooling (see Section 2).

Classification experiments with two medical datasets (ICPR HEp-2 cells and IRMA radiographs) with different types of local features and classifiers showed that the HMP, when combined with any traditional pooling, performs significantly better than the traditional pooling alone. This supports that the proposed HMP captures intermediate structure information which has not been encoded in traditional pooling. Experiments also show that the HMP performs better than region-based pooling for medical data in which there is no meaningful left-right or top-down information.

2. HIERARCHICAL MIX-POOLING

Let $\mathbf{X} = \{\mathbf{x}_i, i = 1 \dots N\}$ be a set of $d-$dimensional local features ($\mathbf{x}_i \in \mathbb{R}^d$) extracted from an image $\mathbf{I}$, and matrix $\mathbf{D} = [\mathbf{w}_1, \mathbf{w}_2, \dots \mathbf{w}_M]$ denote the dictionary of M visual words $\mathbf{w}_i \in \mathbb{R}^d$. At the coding stage of SC approaches, every local feature $\mathbf{x}_i$ can be represented by a sparse linear combination of visual words from the dictionary. The coefficient vector $\boldsymbol{\alpha}_i \in \mathbb{R}^M$ of the combination is a sparse vector representing the local feature $\mathbf{x}_i$ and obtained by minimizing a $L1$-norm regularization problem [1], i.e.,

$$\alpha_i = \underset{\mathbf{a}_i}{\operatorname{argmin}} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda \|\mathbf{a}_i\|_1 \,. \tag{1}$$

Where λ is a regularization parameter. In the traditional pooling stage, a pooling operator g is applied to all the sparse vectors to generate the feature representation $\mathbf{y}$ for image $\mathbf{I}$, i.e.,

$$\mathbf{y}(\mathbf{I}) = g(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_N)\,. \tag{2}$$

Two traditional *pooling* operators, *sum pooling* and *max pooling*, are normally used as defined below [12, 4],

$$\text{sum pooling}: \quad y_s^j(\mathbf{I}) = \sum_{i \in \mathcal{N}(\mathbf{I})} \alpha_{ij} \tag{3}$$

$$\text{max pooling}: \quad y_m^j(\mathbf{I}) = \max_{i \in \mathcal{N}(\mathbf{I})} \alpha_{ij}\,. \tag{4}$$

where y_s^j and y_m^j are respectively the j^{th} element of the pooling results $\mathbf{y}_s$ and $\mathbf{y}_m$, and $\mathcal{N}(\mathbf{I})$ is the set of indices of local features in the image $\mathbf{I}$. Generally, $\mathbf{y}_m$ is the preferred image representation, associated with better performance than $\mathbf{y}_s$ [13]. It is clear that both sum pooling and max pooling fail to encode intermediate structure information because they operate on sparse vectors for small-size (e.g., 16 × 16 pixels) regions.

To encode intermediate structure information, we propose a two-level pooling. First, an image is divided into overlapping intermediate-size regions $\mathbf{r}$, with each region $\mathbf{r}$ consisting of $S \times S$ local features (see Section 3 for regions size details). Pooling (e.g., sum pooling) of the sparse vectors for the multiple local features in the region $\mathbf{r}$ generates a feature representation which is expected to encode intermediate structure information. Then, a different pooling operator (e.g., max pooling) is applied, not to the local features, but to the intermediate-structure representations over all the intermediate-size regions $\mathbf{r}$ to form the final image representation. By using different pooling operators at different levels (therefore *hierarchical mix-pooling*) for the sparse codes associated with the local features and for the intermediate-structure representation, we expect that the final image representation will encode intermediate structure information rather than just encoding the statistics from local features. Also, since the final representation results from a pooling rather than a concatenation of region-based features, it is more invariant to translation of regions of interest in an image. Based on traditional sum and max pooling, we can define two HMP operators:

$$\text{summax pooling:} \quad y_{sm}^j(\mathbf{I}) = \max_{\mathbf{r} \in \mathcal{R}(\mathbf{I})} \sum_{i \in \mathcal{N}(\mathbf{r})} \alpha_{ij}, \tag{5}$$

$$\text{maxsum pooling:} \quad y_{ms}^j(\mathbf{I}) = \sum_{\mathbf{r} \in \mathcal{R}(\mathbf{I})} \max_{i \in \mathcal{N}(\mathbf{r})} \alpha_{ij}\,, \tag{6}$$

where y_{sm}^j and y_{ms}^j are respectively the j^{th} element of the image representation $\mathbf{y}_{sm}$ and $\mathbf{y}_{ms}$, and $\mathcal{R}(\mathbf{I})$ is the set of intermediate-size dense square regions covering the image. As different pooling operators lead to different image statistics [1], we can expect that the proposed HMP representation, capturing intermediate-structure features, adds valuable discriminative information to representations based on local feature information only (Equations 3 and 4).

3. EXPERIMENTS

3.1. Materials and methods

Two public medical image datasets, ICPR cells and IRMA radiographs were used to evaluate the classification performance.

ICPR HEp-2 cell images dataset: This dataset[1] contains $13,596$ gray-scale cell images from 6 different classes (homogeneous, speckled, nucleolar, centromere, golgi, and nuclear membrane), with average image size about 70 × 70. Some images from this dataset are shown in Figure 1.

IRMA radiology dataset The Image Retrieval in Medical Applications (IRMA)-2009 dataset[2] contains 15,363 anonymous radiographs from 57 categories. The images are resized

[1] http://i3a2014.unisa.it/

[2] Thanks to Prof. Thomas M. Deserno, Department of Medical Informatics, University Hospital Aachen, Germany for providing the access to the IRMA dataset. https://ganymed.imib.rwth-aachen.de/irma/index_en.php

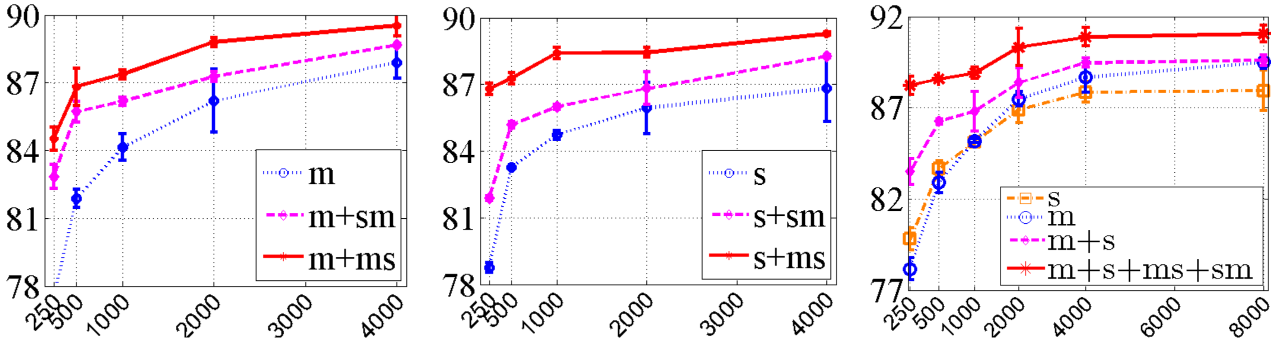


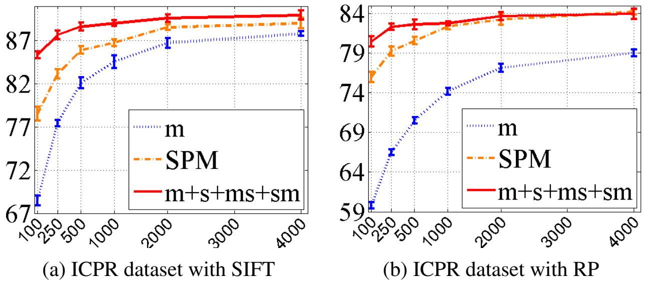
Fig. 2: Effect of HMP on the ICPR HEp2 cell images dataset (SIFT features and SVM classifier). The curves show the MCA (vertical axis) for different dictionary sizes (horizontal axis). The vertical intervals in the curves show the corresponding standard deviations. 'm', 's', 'ms' and 'sm' represent the max (Eqn. (4)), sum (Eqn. (3)), maxsum (Eqn. (6)) and summax (Eqn. (5) poolings respectively.

to be no larger than 300 × 300 with preserved aspect ratio. Since the number of images is very unbalanced across classes, only 20 classes were selected, each of which contains 200 images.

For both datasets, we applied two-fold cross-validation (iterated 5 times) and report the mean per-class accuracies (MCA). Three local features, SIFT, raw patches (RP) and intensity histograms (IH) were used. All these were densely extracted from patches of size 16 × 16 pixels with an overlap of 12 pixels along both horizontal and vertical directions. Locality-constrained linear coding (LLC) [12], an efficient variant of SC was used for feature encoding. The number of nearest neighbours in LLC was set to 10. One-vs-rest multiclass SVM with linear kernel [14] as well as k-NN classifiers were used. SVM parameters were learned using 5-fold cross-validation on the training set. For k-NN, 5 nearest neighbours were experimentally adopted. For the size of intermediate-size square regions $\mathbf{r}$ which consist of $S \times S$ local features, S was experimentally set to 5, and 12 for ICPR and IRMA respectively. In all the experiments, $L2$ normalization is applied to features obtained by each pooling operator.

3.2. HMP vs traditional pooling

To show the effect of HMP, the classification performance was compared with different features with traditional pooling and with HMP, e.g., $\mathbf{y}_m$ vs. $[\mathbf{y}_m^{\mathrm{T}}, \mathbf{y}_{sm}^{\mathrm{T}}]^{\mathrm{T}}$. Figures 2(a)(b) show that the combination of one HMP with one traditional pooling performs significantly better than the traditional pooling alone, achieving competitive accuracies with smaller dictionary sizes. For example, in Figure 2(a), 'm+ms' pooling with dictionary size 500 achieves even better performance than max pooling with dictionary size 2000 (87.3% vs. 86.2%). Figure 2(a) also shows that even after counterbalancing the effect of feature dimensionality, e.g., with the same dimension size 1000, the classification accuracy based on combination of HMP (especially the maxsum pooling) with traditional pooling is significantly higher than the traditional pooling alone (e.g., 87.3% vs. 84.1% in Figure 2(a)). Furthermore, we observed that the combination of all HMP and traditional pooling (e.g., 'm+s+ms+sm' in Figure 2(c)) performs significantly better than any other combinations (e.g., 'm+s' in Figure 2(c) or 'm+ms' in Figure 2(a)). Therefore in the following tests, the whole combination will be used.



(a) ICPR dataset with SIFT

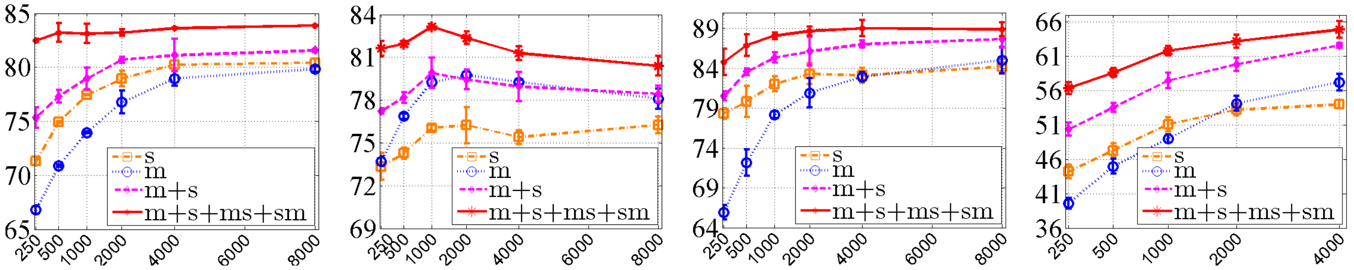
(b) ICPR dataset with RP

Fig. 3: Comparison of HMP with SPM.

3.3. Comparison with region-based pooling

This experiment compares HMP with region-based pooling on the ICPR dataset with two features, SIFT and RP. In the interest of a fair comparison we keep similar feature dimensionality for both pooling methods. For region-based pooling, a two-level SPM is used, where max-pooling over the whole image as well as (2 × 2) image regions are concatenated as the image representation.

Figure 3(a) shows that, for SIFT, the combination of traditional pooling with HMP ('m+s+ms+sm') not only outperforms SPM but also gives better accuracy with a lower dimensional feature representation. For example, when the dictionary size is 1000, HMP gives an accuracy of 89% with a feature dimensionality of 4000, but SPM gives 86.8% with



(a) ICPR dataset with RP and SVM (b) ICPR dataset with SIFT and kNN (c) IRMA dataset with RP and SVM (d) IRMA dataset with IH and SVM

Fig. 4: General applicability of HMP demonstrated with different datasets, local features and classifiers. Similar results were obtained for other combinations of local features and datasets, but not shown due to limited space.

dimensionality 5000. Since SPM captures both intermediate-size structure and global-scale spatial layout information, the superior performance of HMP may be attributed to its translation-invariant property, especially considering there is no meaningful translation information (e.g., left vs. right) in the ICPR cell images. Similar results were obtained for RP features (Figure 3(b)), although the performance of both pooling methods converges for larger-size dictionaries.

3.4. General applicability of HMP

The proposed HMP is a quite general pooling method and its performance is independent of the types of local features and classifiers and not limited to any specific dataset. Figure 4 show that, for both medical datasets and two other local features (RP and IH), similar findings were observed, i.e., the combination of HMP with traditional pooling once again performs significantly better than the traditional pooling alone. With a different classifier (i.e., k-NN), Figure 4(b) again shows the consistent superior performance of HMP when combined with traditional pooling. The decrease of accuracy with larger dictionary size can be described to the less discriminative power of the Euclidean distance used for k-NN in higher dimensional space.

4. CONCLUSIONS

We proposed a simple but effective pooling approach, HMP, to capture the statistics of intermediate structures present in images, and to produce a translation invariant image representation. The effectiveness and general applicability of HMP is confirmed by extensive experiments on two public datasets using different local features and classifiers. Our results show that HMP can significantly improve multi-class classification performance when combined with traditional pooling approaches, where the statistics of the local features were captured. Since in this work we focussed on a two-level pooling, future work will experiment with multi-level (> 2) architectures, and combine other representations (e.g. SPM) with HMP.

Acknowledgements This work was partially funded by 2011-2016 EU FP7 ERC project “CODIR: colonic disease investigation by robotic hydrocolonoscopy”, collaborative between the Universities of Dundee (PI Prof Sir A Cuschieri) and Leeds (PI Prof A Neville).

5. REFERENCES

[1] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in *CVPR*, 2009.

[2] S. Manivannan, Ruixuan Wang, E. Trucco, and A. Hood, “Automatic normal-abnormal video frame classification for colonoscopy,” in *ISBI*, 2013.

[3] Jiajia Luo, Wei Wang, and Hairong Qi, “Group sparsity and geometry constrained dictionary learning for action recognition from depth maps,” in *ICCV*, 2013.

[4] Y.-L. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun, “Ask the locals: Multi-way local pooling for image recognition,” in *ICCV*, 2011.

[5] Lingqiao Liu, Lei Wang, and Xinwang Liu, “In defense of soft-assignment coding,” in *ICCV*, 2011.

[6] Yangqing Jia, Chang Huang, and Trevor Darrell, “Beyond spatial pyramids: Receptive field learning for pooled image features.,” in *CVPR*, 2012.

[7] Jimei Yang and Ming-Hsuan Yang, “Learning hierarchical image representation with sparsity, saliency and locality,” in *BMVC*, 2011.

[8] Liujuan Cao, Rongrong Ji, Yue Gao, Yi Yang, and Qi Tian, “Weakly supervised sparse coding with geometric consistency pooling,” in *CVPR*, 2012.

[9] Dingjun Yu, Hanli Wang, Peiqiu Chen, and Zhihua Wei, “Mixed pooling for convolutional neural networks,” in *Rough Sets and Knowledge Technology*, vol. 8818 of *Lecture Notes in Computer Science*, pp. 364–375. Springer International Publishing, 2014.

[10] L. Bo, X. Ren, and D. Fox, “Multipath sparse coding using hierarchical matching pursuit,” in *CVPR*, 2013.

[11] Chaoqun Weng, Hongxing Wang, and Junsong Yuan, “Hierarchical sparse coding based on spatial pooling and multi-feature fusion,” in *ICME*, 2013.

[12] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, “Locality-constrained linear coding for image classification,” in *CVPR*, 2010.

[13] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce, “Learning mid-level features for recognition,” in *CVPR*, 2010.

[14] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, “LIBLINEAR: A library for large linear classification,” *JMLR*, 2008.