SCIENCE CHINA
Information Sciences

• HIGHLIGHT •

January 2018, Vol. 61 018104:1–018104:3
doi: 10.1007/s11432-017-9158-x

Two-stage local constrained sparse coding for fine-grained visual categorization

Lihua GUO[1*], Chenggang GUO[1], Lei LI[1], Qinghua HUANG[1,2*], Yanshan LI[2] & Xuelong LI[3]

[1]*School of Electronic and Information Engineering, South China University of Technology, Guangzhou* 510641*, China;*
[2]*College of Information Engineering, Shenzhen University, Shenzhen* 518060*, China;*
[3]*The Center for OPTicalIMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an* 710119*, China*

Received 12 February 2017/Revised 25 April 2017/Accepted 21 June 2017/Published online 13 September 2017

Citation Guo L H, Guo C G, Li L, et al. Two-stage local constrained sparse coding for fine-grained visual categorization. Sci China Inf Sci, 2018, 61(1): 018104, doi: 10.1007/s11432-017-9158-x

Fine-grained visual categorization (FGVC) [1, 2] classifies the objects into categories which are both visually and semantically similar. Currently, the FGVC has included many species, i.e., flowers, birds, and dogs. In these species, all categories have a characteristic with intra-class diversity and inter-class similarity.

We propose a two-stage learning framework. In our framework, we use a sparse coding method with local constraint to stably extract the image descriptor. Meanwhile, we design a two-stage learning structure to learn the spatial relationship among these descriptors. The learning structure is implemented through many pathways on multiple patches with varied sizes. Since the part alignment is a necessary preprocess step when objects meet large variation in pose and view, a robust head pose alignment method is designed to eliminate the bad effect from the severe variation in object pose and view point, especially for the bird species. In summary, the main advantages of our work are below: (1) Two-stage learning architecture can capture the invariance in different scales and global spatial relationship in images. (2) Local orientation information is more robustness than the raw pixels. Therefore, our framework can extract more efficient features from local orientation by using local spatial coding. (3) Robust head pose alignment method can guarantee the system to adapt different variations in pose and view.

The system framework. Our system framework is shown in Figure 1, which mainly includes object alignment and features extraction. For adapting the severe variation in pose and view, position alignment is necessary. In our framework, we assume that each part has a geometric relationship, which is a prior information from practical observation. We can calculate the statistical information of each part in all birds' images, and model the part's relationship for automatically locating their positions. Therefore, our pose alignment can be easily extended to the species with a geometric prior among the parts. For the bird species, the part regions of bird's head are first aligned by our automatic position estimation method. Then, a two-stage locally constraint sparse coding architecture is used to extract discriminative features.

In our two-stage framework, the stages I and II are paralleled. Stage I only applies the single layer sparse coding architecture, and stage II

* Corresponding author (email: guolihua@scut.edu.cn, qhhuang@scut.edu.cn)
The authors declare that they have no conflict of interest.

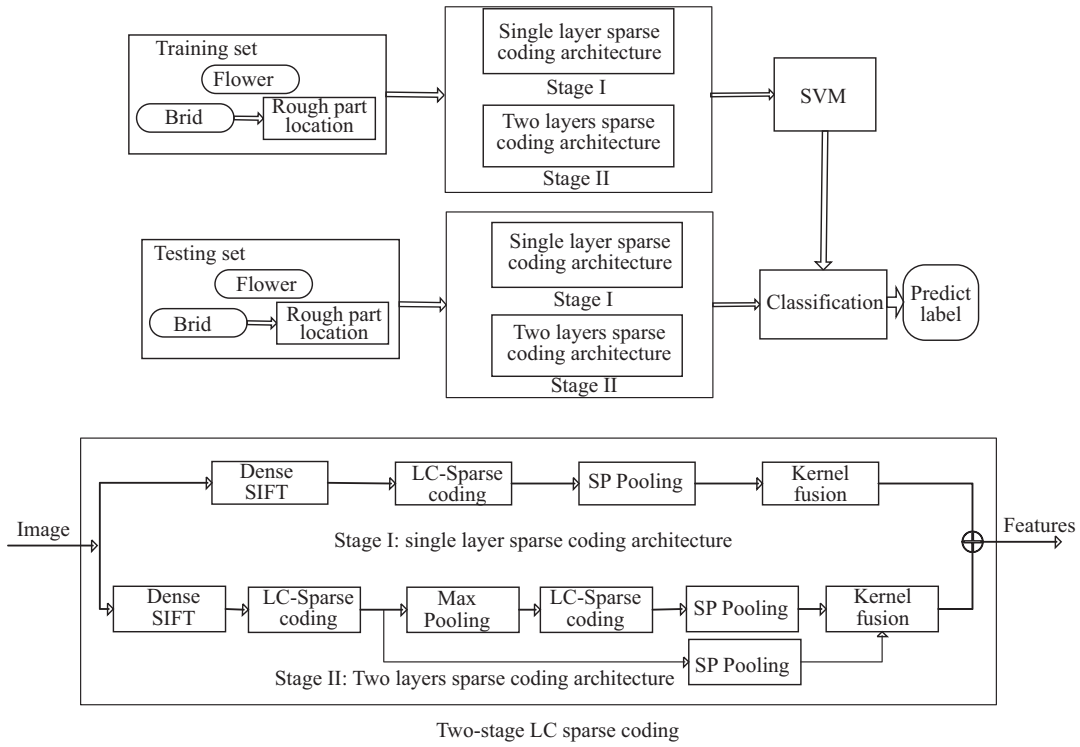

Figure 1 Our proposed framework for fine-grained categorization.

applies two layers sparse coding architecture. In stage I, the system extracts a dense SIFT (scale-invariant feature transform) from the image block, then encodes it using a locally constraint sparse coding, further uses Spatial Pyramid (SP) Pooling to model the feature's spatial relationship. The kernel function is finally used to fuse features. In stage II, a two-layer sparse coding structure is designed, where the first layer is cascaded with the second layer. Outputs of every layer are incorporated by a SP Pooling to model the feature's spatial relationship. Finally, the system uses a kernel to combine all features from each layer, and feeds these features into a SVM classifier.

Automatic pose alignment. In the birds dataset [2], the object pose and view point have a large variation. From our observation, four parts, e.g., eye, beak, forehead and crown, are always visible, and they have some geometric relationship. In the Caltech 200 Birds dataset, the coordinates of four parts are given, and they are labeled in all training images. If all coordinates of four parts in our dataset are directly mapped into a 2D space, their distributions have no regular pattern. However, if we normalize the bird's head directions as the same direction, e.g., heading right, the distribution of four parts coordinates looks like Gaussian distribution. If we only give an approximate proposal to cover the part region instead of accurately detecting the part location, head alignment becomes possible in this study. The main procedure is as follows:

(a) The direction of the bird's head must be previously predicted in the testing images, and the images heading left are flipped into heading right. In the Caltech 200 Birds dataset, the training images are therefore manually classified into three head directions. For judging the head direction, we calculate the histogram of gradient (HOG) value of the bird's head from the testing image, and compare its HOG value with those from all training images. The head direction of the test image can be obtained by voting of three images with the most similar HOG value. If the head direction of test image is left, then it will be mirror mapped and become heading right.

(b) Statistical location information in training images transfers into testing images to achieve an approximate proposal to cover the four parts. The statistical distribution can be calculated from all training images. First, coordinates of four head parts are normalized. Then, the histogram of the four parts coordinate in all training images is calculated. Finally, we fit Gaussian probability density to match the statistical data of each part.

Feature extraction. Sparse coding is to represent signals $Y = [y_1, y_2, \ldots, y_N] \in \mathbb{R}^{H\times N}$ as a few nonzero entries $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{M\times N}$ from a prebuilt codebook $D = [d_1, d_2, \ldots, d_M] \in \mathbb{R}^{H\times M}$, which is much redundant as possible in

order to sparsely decompose the sampled signals. One standard optimization approach is to minimize the following reconstruction error by forcing codes to be K sparse level:

$$\begin{aligned} &\min_{D,X} \|Y - DX\|_F^2 \\ &\text{s.t.} \quad \forall m, \|d_m\|_2 = 1, \forall n, \|x_n\|_0 \leqslant K. \end{aligned} \tag{1}$$

M-HMP [3] gave high probability to some frequently observed patches when learning codebooks. Hence the authors added a regularization term for avoiding the over-fitting problem. However, when using M-HMP method, some similar image patches will select different codebooks. It indicates that noises will make the dictionary learning unstable in M-HMP method. Inspired by Locality-Constrained coding (LLC) method [4, 5], we add a local contrained regularization into the M-HMP object function, and our optimization function is as follows:

$$\begin{aligned} &\min_{D,X} \|Y - DX\|_F^2 + \lambda \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} |d_i^{\mathrm{T}} d_j| \\ &\quad + \beta \sum_{i=1}^{N} \|e_i \odot x_i\|^2 \\ &\text{s.t.} \quad \forall m, \|d_m\|_2 = 1, \forall n, 1^{\mathrm{T}} x_n = 1, \end{aligned} \tag{2}$$

where $\odot$ denotes the element-wise multiplication, $e_i = \exp([\mathrm{dist}(y_i, d_1), \ldots, \mathrm{dist}(y_i, d_M)]^{\mathrm{T}}/\sigma)$, $\mathrm{dist}(y_i, d_j)$ is the Euclidean distance between y_i and d_j, and σ is a weight adaptor. The iterative optimization is used to solve (2).

Conclusion. A two-stage locally constraint sparse coding framework is proposed to solve the FGVC. This two-stage framework is used to learn intermediate-level features, and the locally constraint term keeps dictionary learning stable. The local orientation histogram takes the place of raw pixels for extracting more discriminative information during sparse coding. Therefore, the dictionary updating process quickly converges after small iterative times. Moreover, a pose estimation is proposed to make region alignment for eliminating the bad effect from the severe variation in object pose and view point.

Acknowledgements This work was partially supported by National Natural Science Foundation of China (Grant Nos. 61372007, 61571193), Natural Science Foundation of Guangdong Province (Grant No. 2015A030313210), Guangdong Science and Technology Program (Grant No. 2017A010101027), Guangzhou Science and Technology Program (Grant Nos. 201605130119420, 201707010141), and Fundamental Research Funds for the Central Universities (Grant No. 2015ZM138).

References

1 Nilsback M E, Zisserman A. Automated flower classification over a large number of classes. In: Proceedings of Indian Conference on Computer Vision, Graphics and Image Processing, Madurai, 2008. 2: 722–729
2 Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 Dataset. Computation and Neural Systems. Technical Report, CNS-TR-2011-001. 2011
3 Bo L F, Ren X F, Fox D. Multipath sparse coding using hierarchical matching pursuit. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, 2013. 2: 660–667
4 Guo L H. Locality-constrained multi-task joint sparse representation for image classification. IEICE Trans Inform Syst, 2013, 96: 2177–2181
5 Wang J J, Yang J C, Yu K, et al. Locality-constrained linear coding for image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 2010. 9: 3360–3367