# Hierarchical Part Matching for Fine-Grained Image Classification

Lingxi Xie[1], Qi Tian[2], Richang Hong[3], Shuicheng Yan[4] and Bo Zhang[1]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Department of Computer Science, University of Texas at San Antonio, Texas, USA

## ABSTRACT

As a special topic in computer vision, fine-grained visual categorization (FGVC) has been attracting growing attention these years. Different with traditional image classification tasks in which objects have large inter-class variation, the visual concepts in the fine-grained datasets, such as hundreds of bird species, often have very similar semantics. Due to the large inter-class similarity, it is very difficult to classify the objects without locating really discriminative features, therefore it becomes more important for the algorithm to make full use of the part information in order to train a robust model.

In this paper, we propose a powerful flowchart named Hierarchical Part Matching (HPM) to cope with fine-grained classification tasks. We extend the Bag-of-Features (BoF) model by introducing several novel modules to integrate into image representation, including foreground inference and segmentation, Hierarchical Structure Learning (HSL), and Geometric Phrase Pooling (GPP). We verify in experiments that our algorithm achieves the state-of-the-art classification accuracy in the Caltech-UCSD-Birds-200-2011 dataset by making full use of the ground-truth part annotations.

## NOVELTY

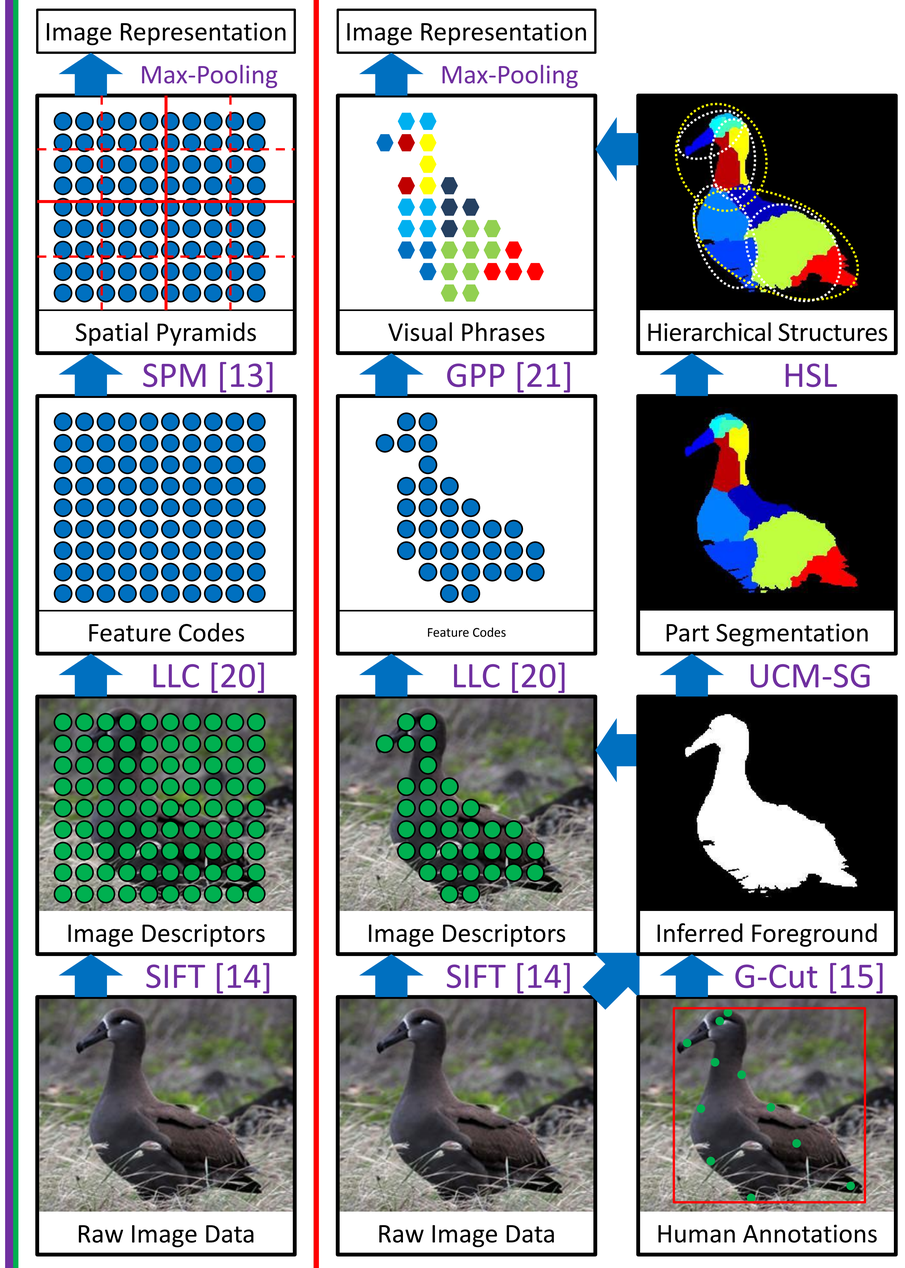
In this paper, we propose a novel flowchart named Hierarchical Part Matching (HPM) to cope with fine-grained classification problems. We make full use of the ground-truth part annotation to help us obtain better image alignment and segmentation, and provide a much more descriptive image representation by building mid-level structures on local features as well as segmented regions.

1. We adopt the Grab-Cut algorithm [15] and the Ultrametric Contour Map [1] on the ground-truth annotations to infer the object (foreground) on the image and segment it into semantic parts.
2. We propose the Hierarchical Structure Learning (HSL) algorithm to find mid-level concepts beyond basic parts.
3. We use the Geometric Phrase Pooling (GPP) algorithm to capture mid-level structures in the local feature groups.

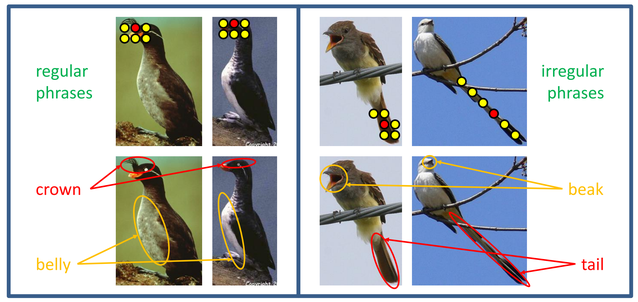
Integrating all the modules above gives a powerful model, which achieves the state-of-the-art classification performance in a challenging fine-grained image collection.

The success of our algorithm encourages the computer vision society to design more effective part detection methods to help the fine-grained object recognition task.

## THE PROPOSED FRAMEWORK



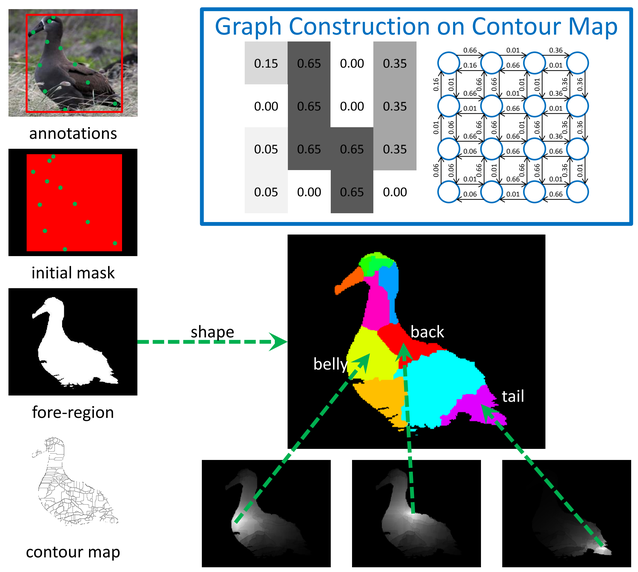

### GPP helps capturing geometric contexts




### HSL helps constructing mid-level concepts

| No. | μ | Learned Mid-level Parts |
|---|---|---|
| #1 | 0.0 | No mid-level parts are learned. |
| #2 | 0.1 | **eyes** (left/right eye), **legs** (left/right leg), **wings** (left/right wing). |
| #3 | 0.3 | **eyes**, **legs**, **wings**, **neck** (nape/throat). |
| #4 | 0.5 | **eyes**, **legs**, **wings**, **neck**, **head** (beak/crown/forehead/**eyes**), **body** (back/belly/breast/tail). |
| #5 | 1.0 | **eyes**, **legs**, **wings**, **neck**, **head**, **body**, (**wings**/**legs**), (**body**/**wings**/**legs**), ALL. |

### Foreground Inference and Segmentation




## RESULTS

### Random Splits

| #training | 5 | 10 | 20 | 30 | 40 |
|---|---|---|---|---|---|
| Wah [18] | 10.05 | | | | |
| Wang [19] | 13.64 | 20.25 | 28.36 | 33.63 | 37.77 |
| Xie [20] | 15.34 | 22.91 | 31.01 | 36.17 | 40.43 |
| Ours | 36.09 | 48.87 | 60.56 | 65.62 | 69.07 |

### Fixed Splits (~30 trains)

| Wah [18] | Wang [19] | Xie [20] | Zhang [24] | Ours |
|---|---|---|---|---|
| 17.31 | 33.91 | 36.33 | 24.21 | 66.35 |

### Conclusions

We propose a novel framework for fine-grained image classification, and evaluate it on the challenging CUB-200-2011 dataset. We add three modules into the BoF model to make full use of the ground-truth landmark annotations. The integrated algorithm outperforms the state-of-the-art competitors significantly.

## REFERENCES

Key references are numbered as they appear in the paper.
[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. CVPR, 2006.
[14] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. IJCV, 2004.
[15] C. Rother, V. Kolmogorov, and A. Blake. GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts. ACM ToG, 2004.
[18] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. TechRep, 2011.
[20] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-Constrained Linear Coding for Image Classification. CVPR, 2010.
[21] L. Xie, Q. Tian, and B. Zhang. Spatial Pooling of Heterogeneous Features for Image Applications. ACM Multimedia, 2012.
[24] N. Zhang, R. Farrell, and T. Darrell. Pose Pooling Kernels for Sub-Category Recognition. CVPR, 2012.

## ACKNOWLEDGE.

This work was supported by the National Basic Research Program (973 Program) of China under Grant 2012CB316301, and Basic Research Foundation of Tsinghua National Laboratory for Information Science and Technology (TNList). This work was also supported in part to Dr. Qi Tian by ARO grant W911NF-12-1-0057, NSF IIS 1052851, Faculty Research Awards by Google, NEC Laboratories of America and FXPAL, UTSA START-R award and NSFC 61128007, respectively. This work was also supported by the Singapore National Research Foundation under its International Research Centre at Singapore Funding Initiative and administered by the IDM Programme Office.