Master's Thesis

Recognition of Flower Species Using Visual Vocabulary of Compound Descriptors

Department of Electronics and Computer Engineering
Graduate School, Chonnam National University

Aich Shubhra

August 2016

Recognition of Flower Species Using Visual Vocabulary of Compound Descriptors

Department of Electronics and Computer Engineering
Graduate School, Chonnam National University

Aich Shubhra

Supervised by Professor Lee, Chil-Woo

A dissertation submitted in partial fulfillment of the requirements for the Master of Science in Electronics and Computer Engineering

Committee in Charge :

Park, Jaehyung ____________

Kim, Cheol-Hong ____________

Lee, Chil-Woo ____________

August 2016

Table of Contents

Table of Contents ···································· i

List of Tables ···································· ii

List of Figures ···································· iii

Abstract ···································· iv

1. Introduction ···································· 1

2. Related Work ···································· 4

3. Our Approach ···································· 7

3.1. Segmentation ···································· 7

3.2. Composite Descriptors ···································· 10

3.3. Classifier ···································· 18

4. Experiment ···································· 20

4.1. Dataset ···································· 20

4.2. Results ···································· 21

5. Conclusion ···································· 24

Reference ···································· 26

Abstract (Korean) ···································· 33

<List of Tables>

[Table 4.1] Comparison Between Conventional and Composite Descriptors 22

[Table 4.2] Recognition Performance of Features on Oxford-102 Dataset 23

[Table 4.3] Performance Comparison on Oxford Dataset 24

<List of Figures>

[Figure 1] Sample images from dataset images ······································ 2

[Figure 3.1] Typical pattern classification system ································· 7

[Figure 3.2] Segmentation results ··· 10

[Figure 3.3] Bag-of-visual-words modeling ·· 12

[Figure 3.4] SIFT extraction ·· 13

[Figure 3.4] HOG extraction ·· 17

[Figure 3.5] Block diagram ·· 19

[Figure 4] Distribution of test images ··· 21

Recognition of Flower Species Using Visual Vocabulary of Compound Descriptors

Aich Shubhra

Department of Electronics and Computer Engineering
Graduate School, Chonnam National University,
(Supervised by Professor Lee, Chil-Woo)

(Abstract)

We deal with the problem of flower classification which falls into the broader category of fine-grained recognition. We investigate the classification performance of the bag-of-words models of various descriptors covering different aspects of flower properties. To this end, we propose to apply the well-established descriptors along with the local image statistics in a more elaborate way to derive extra information from RGB flower images taking each color plane into account individually. We demonstrate that the composite descriptors constructed this way outperform their conventional counterparts to a great extent in the domain of flower classification. We evaluate the bag-of-visual words model of the proposed composite descriptors on Oxford-102 class dataset using multiple-kernel learning algorithm as the classifier. Our vocabulary of compound descriptors shows the improvement with the accuracy of 85.02%

on Oxford-102 dataset over other methods using the training set of this dataset for supervision.

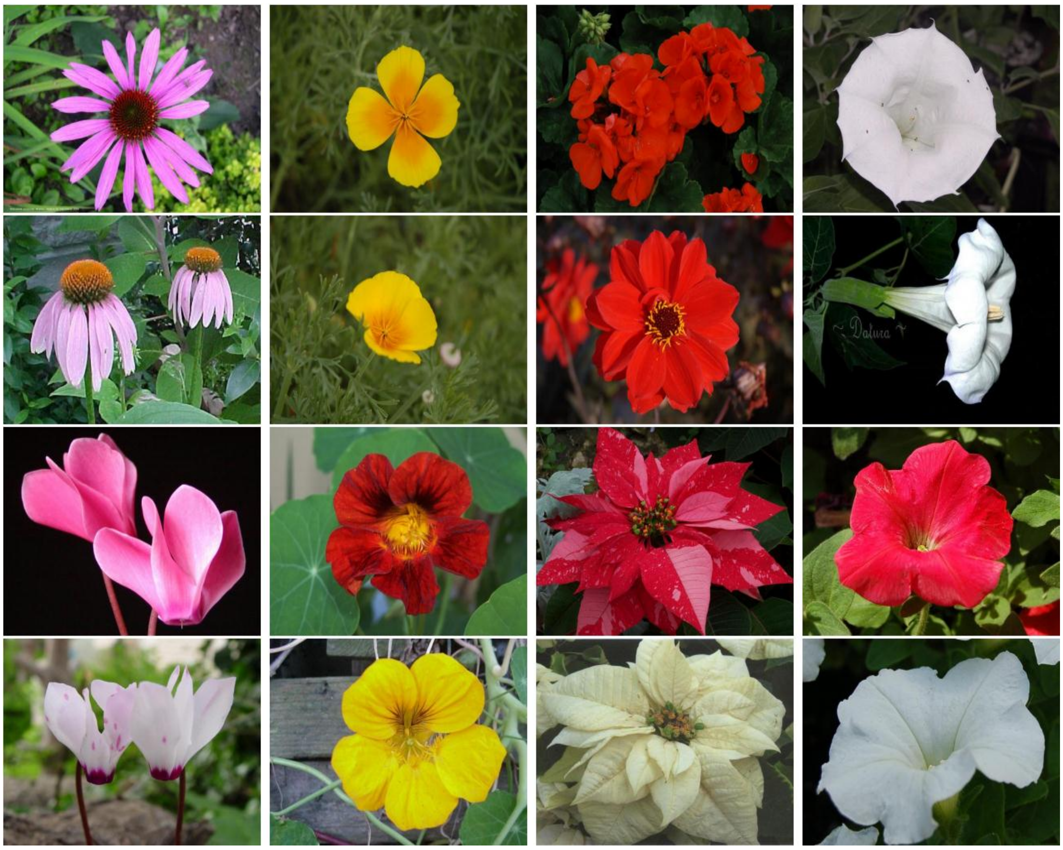
Keywords: Flower classification, compound visual vocabulary, local statistics, SIFT, HOG, multiple kernel learning.

1. Introduction

Object recognition is one of the most fundamental problems in computer vision. Most of the researches in this area belong to the classification problem among different categories, like cars, airplanes, animals, etc. as we can see from PASCAL [1] and Caltech [2] datasets. Rather than differentiating the objects of different categories, we investigate the recognition problem within the single basic category, which is flower. This type of problem is also referred to as fine-grained recognition and in most cases, it requires expert domain knowledge.

The problem of flower classification appears to be more challenging than generalized object classification for several reasons. One reason is large inter-species shape and color similarities. Also, mostly the overall shape information of the flowers is quiet useless because of the high-level of deformation of non-rigid flower petals. Therefore, it is very difficult for the laymen to identify the species of a given flower from a color image with visual inspection. Given only the image, sometimes it becomes even impossible for the botanists or flora-experts to differentiate among visually similar flower categories.

Sample flower images from Oxford-102 dataset [3] are shown in Figure 1. In this figure, first two rows illustrate the fact of within category color similarity and shape dissimilarity. Each column in the first two rows contains images of flowers from same species. Although the colors of the



[Figure 1] Sample images from Oxford 102 flower dataset [3]. First two rows illustrate the fact of within color similarity and shape dissimilarity whereas for the last two rows, it is vice versa. Each column in the first and the last two rows contains images of flowers from same species. For the first two rows, even though the colors of the same species are similar, their shapes are quite different because of viewpoint changes and/or deformability of flower petals. Hence, for all these cases, color based features visually appear to be of significant importance. On the contrary, for the last two rows, albeit the colors of the same species are absolutely different, their intra-class shapes and inherent textures seem to be similar enough for recognition.

same species are similar, their shapes are quite different because of viewpoint changes and / or deformability of flower petals. Hence, for all these cases, color based features are of much importance. On the other hand, the last two rows demonstrate the aspect of intra-class color dissimilarity and shape similarity. Like the first two rows, each column in the last two rows contains images of flowers from same species. Even though the colors of the same species are completely different, their shapes and inherent textures are quite similar. Therefore, in this case, shape and texture based features are more crucial than that of colors. Thus, from this observation, it is evident that features based on solely color or shape or texture properties of flower would be insufficient to achieve notable recognition performance. This hypothesis is also proved by recent studies [3-6] showing that combination of heterogeneous features covering the aspects of texture, shape and color is successful in flower classification since the heterogeneity of the feature sets exploit the subtle object properties complementarily. Hence, in this paper, we use the bag-of-words modeling approach of various feature descriptors, similar to [3, 4].

To this point, our novelty is in applying local statistics and standard feature descriptors in a way to extract more detailed information from the images. The composite descriptors constructed this way excel their conventional counterparts to a great extent for the flower classification problem. We evaluate the performance of our framework on Oxford-102 dataset [3] using the multiple-kernel learning algorithm of Tang et al. [7].

We achieve the best recognition performance amongst all the recent methods using the training set of this dataset only for their supervision.

This paper is organized as follows. Next section gives a brief account of the related works. Our approach is described in section 3. A short description of Oxford-102 dataset and the experimental results are illustrated in section 4. Finally, section 5 provides the conclusion of the paper.

2. Related Work

Very recently, deep learning based frameworks have gained lots of attention due to the tremendous success in different domain of artificial intelligence, e.g. speech recognition and computer vision. This triumph of deep convolutional neural nets has been commenced in the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC 2012). Hinton and his associates [33] build the network which automatically learns useful image representations for the classification task using the pixels as input. The results reveal that deep learning significantly outperforms state-of-the-art computer vision representation competitors to classify millions of high-resolution images from ImageNet dataset [33]. There are a couple of works in the domain of flower classification also [34, 35]. However, although this trend is unquestionable for this large-scale context (around or more than 1 million training examples), the feasibility of reaching state-of-the-art performances in other complex datasets with fewer training examples remains unclear. And so, bag-of-words (BoW) model [22] that appeared as one of the most prominent approaches in the last decade still remains a very competitive representation. Two main milestones boosted the BoW model strength. The first one is the design of discriminative low-level features, such as Scale-Invariant Feature Transform (SIFT) [11], Speeded Up Robust Features (SURF) [37] and Histograms of Oriented Gradients (HOG) [12]. The second one is the emergence of mid-level representations inspired from the text retrieval community. Indeed, coding and pooling of local features provide a way to construct the constant length vectorial representation for each image which can further be utilized by the pattern classification algorithms as a single feature vector. Subsequently, this representation has been used to train powerful statistical learning models, i.e. Support Vector Machine (SVM) [36] and its variants. This kind of approach is used in the recognition of flower species. We briefly address them and other relevant literatures below.

Nilsback and Zisserman [3, 4, 8-10] have the first extensive level of work on flower classification. They use the bag-of-visual words model of combination of HSV, internal and boundary SIFT [11] and HOG [12] with a non-linear multiple kernel support vector machine [13]. Also, later they propose geometric layout features [9] to make all the images of same size and orientation. Our work can be considered a modified extension, improvement and generalization over their work.

Chai et al. [5] proposes two iterative bi-level co-segmentation algorithms (BiCos and BiCos-MT) based on SVM classification using GrabCut [15] and high-dimensional descriptors stacked from the standard sub-descriptors such as, color distribution, SIFT [11], size, location within the image and shape – all extracted from superpixels [16] of single image and multiple images from multiple classes. Finally, they use concatenated bag-of-words histograms of LLC [17] quantized Lab color and three different SIFT descriptors extracted from the foreground region and linear SVM for recognition.

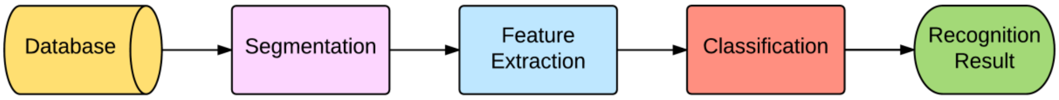
Ito and Kubota [6] use the concept of co-occurrence features for recognition. They propose three heterogeneous co-occurrence features, i.e. color-CoHOG which consists of multiple co-occurrence histograms of oriented gradients including color matching information, CoHED which is the co-occurrence of edge orientation and color difference, CoHD which is the co-occurrence of a pair of color differences and lastly, one homogeneous feature – color histogram [4].

Angelova and Zhu [18] use RGB intensity based pixel affinity for segmentation and max pooling of the LLC [17] encoded HOG features [12] at multiple scales with linear SVM for classification.

Murray and Perronnin [19] propose generalized max pooling (GMP) that obtains the same level of effect as max pooling and applicable not only to conventional bag-of-visual words models, but also to Fisher vector representations. Their method exhibits the best performance over all the previous methods using Oxford-102 dataset for supervision.

3. Our Approach

In a broad sense, we follow the approach of a typical pattern classification system as shown in figure 3.1. Here, we describe each of the blocks of our system separately in each subsection.



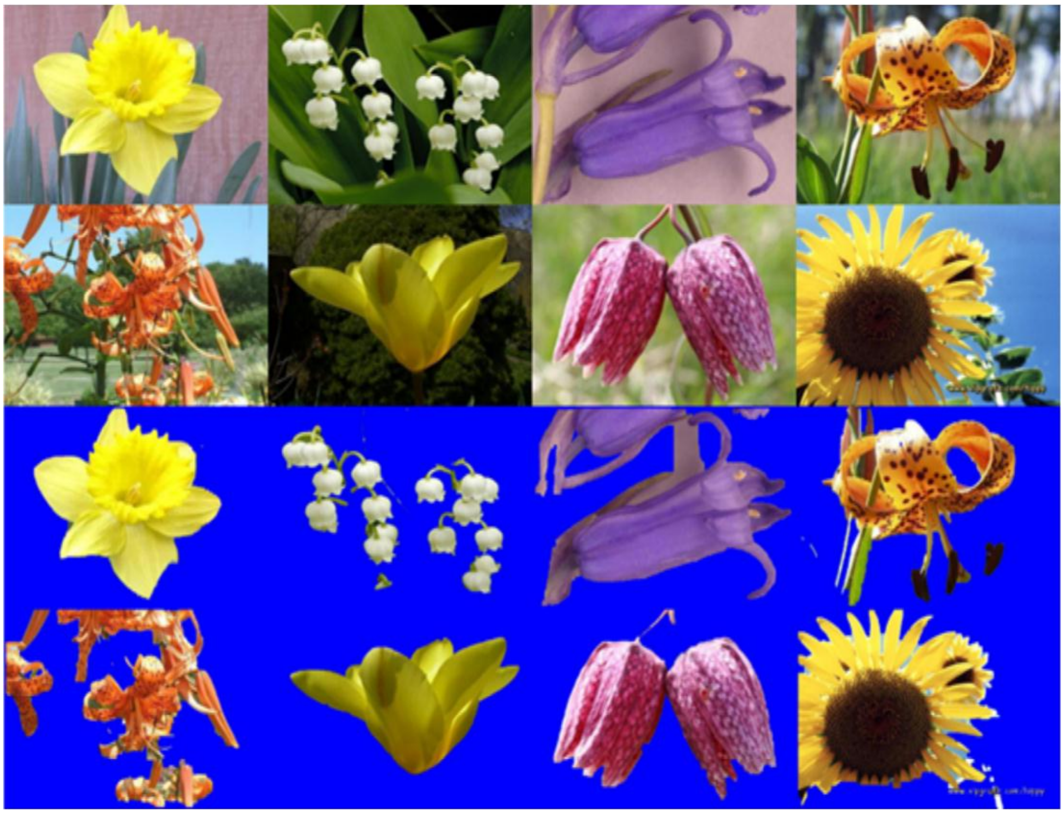
[Figure 3.1] Block diagram of a typical pattern classification system. We leverage the same approach in a broader sense.

3.1 Segmentation

Many object recognition researches leverage segmentation as pre-processing [3, 5, 18, 20]. Our objective in this paper is to demonstrate the strength of the combination of bag-of-visual words models of composite descriptors using multiple kernel learning in the domain of flower classification. So, we use simple GrabCut [15] to segment the flower region from the images, prior to applying our recognition framework. Most segmentation techniques make use of either edge or region information in the image in order to perform segmentation. GrabCut uses both region and boundary information contained in an image to do the segmentation by using graph cuts. This information is used to create an energy function which, when minimized, produces the best segmentation.

GrabCut is basically an interactive algorithm. Initially user draws a rectangle around the foreground region with the condition that the foreground region should be strictly inside the rectangle. Each pixel outside the rectangle is considered in the background for sure whereas the state (foreground / background) of all the pixels inside the rectangle is unknown and is to be determined by the algorithm. Hence, the pixels marked as the background are hard-labeled since there is no uncertainty of being them belonging to the background. Next, a Gaussian mixture model (GMM) is used to model the foreground and background. GMM generates pixel distribution and labels the pixels inside the rectangle as probable foreground or background like clustering. A graph is built from this pixel distribution with pixels as its nodes. Two other nodes are added; source node connected to all the foreground pixels and sink node connected to the background ones. The weights of edges connecting pixels to source and sink nodes are defined as the probability of the pixels being foreground or background. Moreover, the weights of edges connecting the pixels are determined by the pixel similarity based on intensity values. A mincut/maxflow algorithm is used to segment the graph. This algorithm cuts the graph into two separating source and sink node by minimizing the cost function which is the sum of the weights of the cut edges. Finally, pixels connected to source and sink nodes are referred to as the foreground and background pixels, respectively after the cut. All these steps are continued until convergence.

We automate the GrabCut optimization by declaring the bounding box a few pixels inside the image borders with the assumption that the flower object is not located near the borders. Although this assumption is not valid for most of the images, the effect on recognition performance of incorrect segmentation for only few pixels in the border regions is just negligible. Figure 3.2 shows sample images and segmentation results from Oxford-102 dataset.



[Figure 3.2] Sample images from Oxford-102 dataset and their GrabCut segmentations. The first two rows show the sample RGB images and the last two rows show their corresponding segmented images.

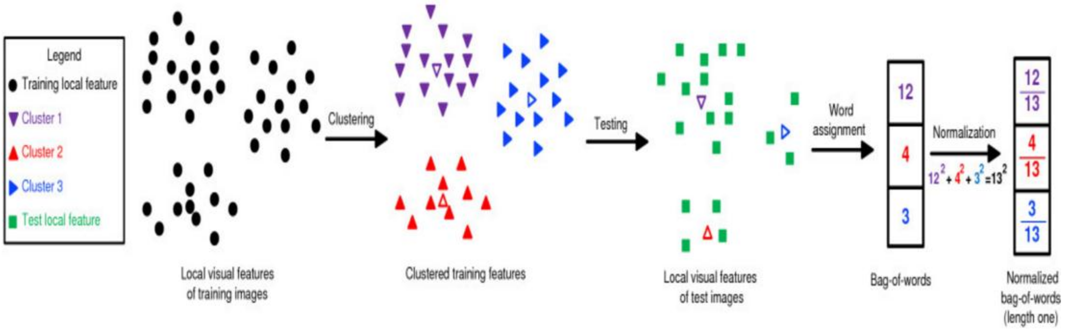
3.2 Composite Descriptors

In this methodology, we consider 4 sets of descriptors – HSV color values with local statistics, 2 variants of SIFT [11] and 1 variant of HOG [12].

Intensity and Local Statistics: In case of object recognition, HSV colorspace is considered for its comparatively less sensitivity to illumination variance [3, 4]. The descriptors are taken as the average of non-overlapping NxN neighborhood over the entire image. Moreover, as the measures of local statistics, we use local standard deviation, range and entropy (equation 1) [21] values of NxN neighborhood for all the three planes of the images. Finally, all 12-D descriptors extracted from the training set are clustered into K disjoint partitions using the K-means algorithm. This kind of clustering of descriptors is known as bag-of-visual words modeling [22]. The values of the parameters N and K are chosen experimentally keeping the bias-variance tradeoff in mind. A pictorial representation of bag-of-visual words is shown in figure 3.3 [29].

$$\begin{aligned}
&\sigma = \sum_{i=0}^{N-1} p(x_i)\,(x_i - \bar{x})^2 \\
&r = \max\{x_i\} - \min\{x_i\} \qquad i = 0, 1, \ldots, N-1 \\
&e = -\sum_{i=0}^{N-1} p(x_i)\log_2 p(x_i)
\end{aligned} \tag{1}$$

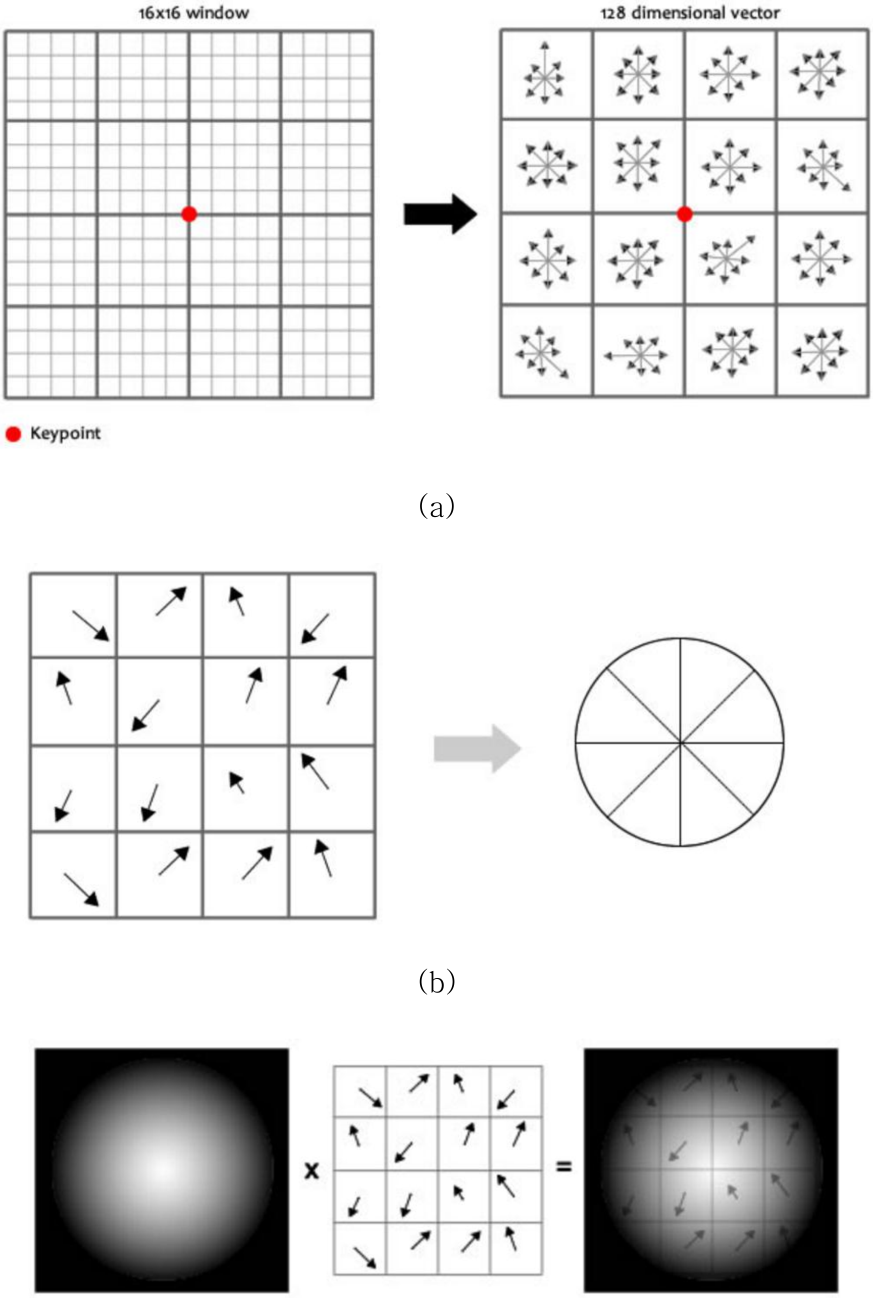
Here, σ, r and e are the local standard deviation, range, and entropy, respectively. N is the number of pixels in the neighborhood, x_i is the intensity of the single color plane for pixel i, $\bar{x}$ is the local mean and $p(x_i)$ is the probability of x_i in the neighborhood.



[Figure 3.3] A pictorial representation of bag-of-visual words modeling [29].

Dense SIFT (D-SIFT): SIFT descriptors take the local shape and texture properties of the objects in the segmented images into account. SIFT [11] is originally proposed based upon a model of biological vision, in particular of complex neurons of primary visual cortex [30, 31]. Most neurons in primary visual cortex are selective to line or gradient orientation and scale or spatial frequency. Some of them are sharply tuned to orientation and fail to respond to lines that are just a bit tilted from their preferred orientation while other cortical neurons are broadly tuned and respond to a broad range of orientations. The steps in SIFT [11] extraction include finding the key points or the points of maxima and minima in the difference of Gaussian (DoG) images calculated from the scale space representation of the original image, removal of the bad key points using a technique similar to Harris corner detector, assigning one or multiple orientations to each of the key points and finally, generate the descriptor vectors from the key points with assigned orientation and scale.

A pictorial representation of generating SIFT descriptors is given in figure 3.4 below.



(a)

(b)

(c)

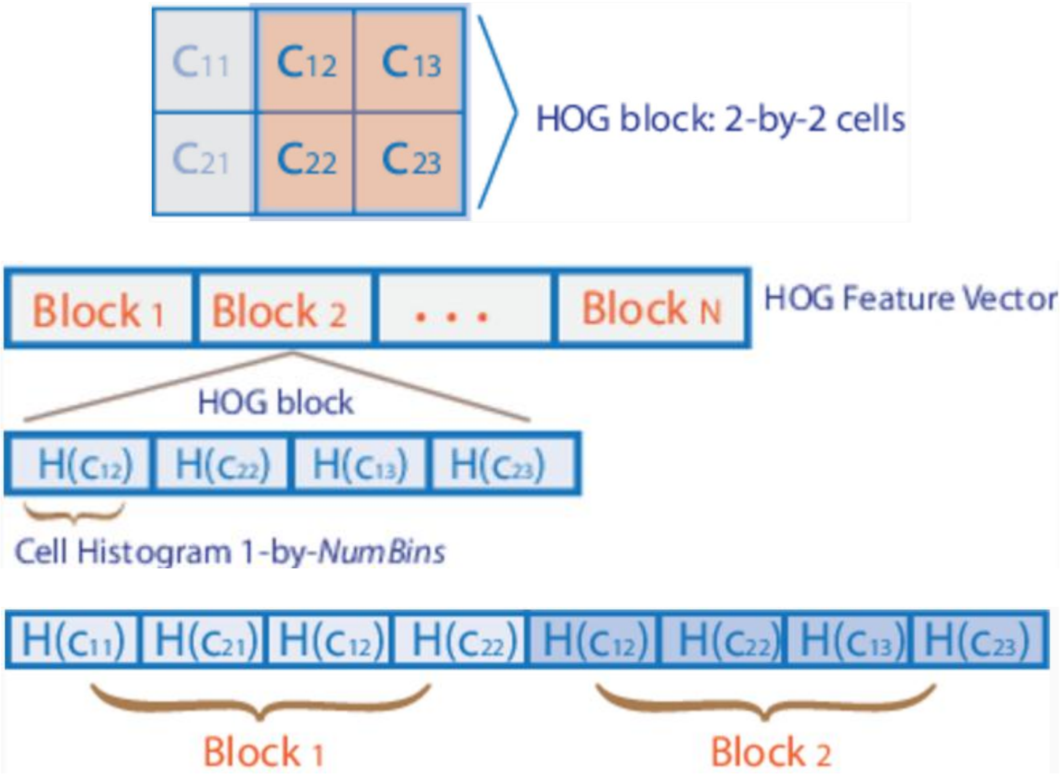
[Figure 3.4] SIFT descriptor extraction [32]: (a) Take a 16x16 window around the keypoint and split this into sixteen 4x4 disjoint windows. (b)

Then, extract 8 bin histograms, one from each 4x4 window and finally cascade them to construct 8x16 = 128 dimensional descriptor vector. (c) The amounts added to the bin for a single pixel depends on the magnitude of the gradient weighted by Gaussian kernel in that pixel.

However, the original scale estimation procedure of Lowe [11] in the SIFT calculation shows serious drawbacks. Objects naturally appear in images in arbitrarily different scales. In most cases, in natural images, these scales are unknown and so multiple scales should be considered for each feature point. One typical approach is to seek for each feature point stable, characteristic scales to both reduce the computational complexity of higher level visual systems, as well as improving their performance by focusing on more relevant information. In the seminal paper of SIFT [11], Lowe followed this approach. In that paper, Only a small number of interest points over characteristic scales centered on the corner structures in the image are selected for descriptor extraction. However, this method produces a small set of interest points over approximately correct scales. According to the paper of Mikolajczyk [23], a scale change of factor 4.4 causes the percent of pixels for which a scale is detected to reduce as little as 38% for the difference of Gaussian (DoG) operator used in [11]. And among the scale of the detected points, only about 10.6% scale estimation is correct. The limitation behind this scale approximation can be subdued with the regular point sampling strategy [24] by extracting the SIFT descriptors on a regular grid of spacing M with circular patches of fixed radius R [3, 14]. This version of SIFT is also known as Dense SIFT [25] and the reason behind the success of this variant over the original one is having descriptors for many pixels over accurate scales than just having a few with the Lowe's method. In addition, we include SIFT descriptors from red, green and blue planes independently (figure 3.5c) whereas trivial approach to SIFT extraction uses the grayscale images. The rationale behind this non-trivial approach is to retrieve as much discriminatory information from the images as possible and it is also justified by the performance improvement demonstrated in the results section. Finally, all the descriptor points together are used for bag-of-features modeling.

Multi-Scale Dense SIFT (MSD-SIFT): As already stated in the last paragraph, objects appear at arbitrarily different scales in natural images. Dense SIFT fails to handle larger scale differences. A plausible solution for this problem is to extract features over multiple scales for each interest point. Hence, in addition to dense SIFT, we add a variant of multi-scale dense SIFT; pyramid histogram of visual words (PHOW) [14, 26]. To calculate the PHOW descriptor, SIFT is computed over 4 circular patches individually with a pre-specified radii at each point on a regular grid. Unlike the conventional SIFT extraction on RGB to grayscale converted images, first we use all the three planes as separate grayscale images to extract descriptors and then cascade the corresponding descriptors of each point in order to construct higher dimensional descriptors (figure 3.5b). The reason behind such cascading is deducing more detailed information about the image on the interest points and consequently, leading to the performance boost in terms of recognition accuracy. However, we have also tried to combine the descriptors from all three planes using simple element-wise operations to address the problems associated with the curse of dimensionality, but such approaches ultimately result in reducing the discriminatory property of the descriptors to a considerable extent. At last, all the descriptors are clustered into K visual words.

Non-Overlapping HOG (NO-HOG): The theme of HOG feature is to characterize local texture and shape properties of the objects by representing the local intensity gradients in a distributed manner regardless of accurate positional information. HOG provides a local image representation with a handle to control the degree of invariance to local geometric and photometric transformations in terms of the number of spatial and orientation bins. In practice, HOG is implemented [12] as the normalized histograms of oriented gradients with overlaps between the neighboring cells in a grid. Cells are small blocks of pixels. The number of bins B is equal to the number of orientations in the histogram. The basic of HOG feature extraction is displayed in figure 3.5 below.



[Figure 3.5] Overlapping of the histograms between the neighboring cells in HOG calculation [12].

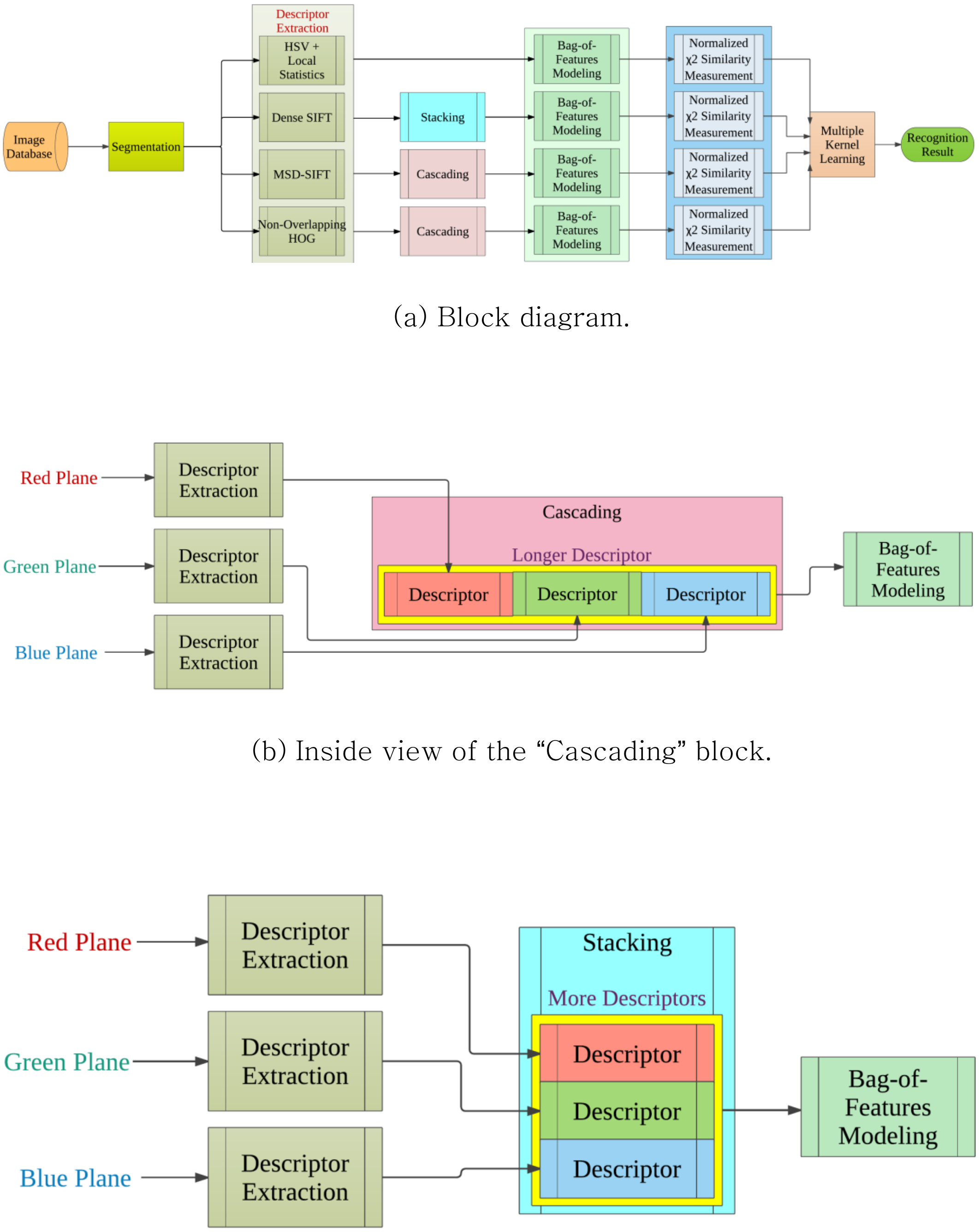
From this figure, we can see that in the original formulation of HOG, overlapping between the neighboring cells occurs resulting in the repetition of the histogram of overlapped cells in the final feature vector. However, like D-SIFT descriptors, we treat B-bin normalized histograms as B-dimensional descriptors. In this modified structure, overlapping appears to be useless by introducing multiple copies of the same B-bin histogram. Therefore, in our calculation, we use zero overlapping between the cells [14]. Like multi-scale SIFT described in the last paragraph, we extract the descriptors from each of the color planes separately and cascade them to form larger descriptors (figure 3.5b) followed by bag-of-feature modeling.

3.3 Classifier

Now, with all these hand-crafted, delicate features at hand, we need a classifier that can deal with the issue of feature-level fusion by itself. In general, multiple kernel learning algorithms serve well for this purpose. This algorithm accomplishes the fusion task by optimizing over the available training set. It uses support vector machine (SVM) as its base classifier and a weighted linear combination of kernels, each kernel corresponding to one feature. Therefore, we utilize the multiple kernel learning approach of Tang et al. [7]. We use normalized chi-square distance (equation 2) [14, 27] to calculate the similarity matrices and same kernel for each feature. Semi-definite linear program (SILP) [28] is used in [7] for the purpose of large scale kernel learning. The block diagram of our approach is shown in figure 3.5 in details.

$$\chi^2(h_1, h_2) = \frac{1}{2}\sum_i \frac{(h_1(i) - h_2(i))^2}{h_1(i) + h_2(i)} \tag{2}$$

Here, h_1 and h_2 are the histograms and i indicates the bin index.



(a) Block diagram.

(b) Inside view of the “Cascading” block.

(c) Inside view of the “Stacking” block.

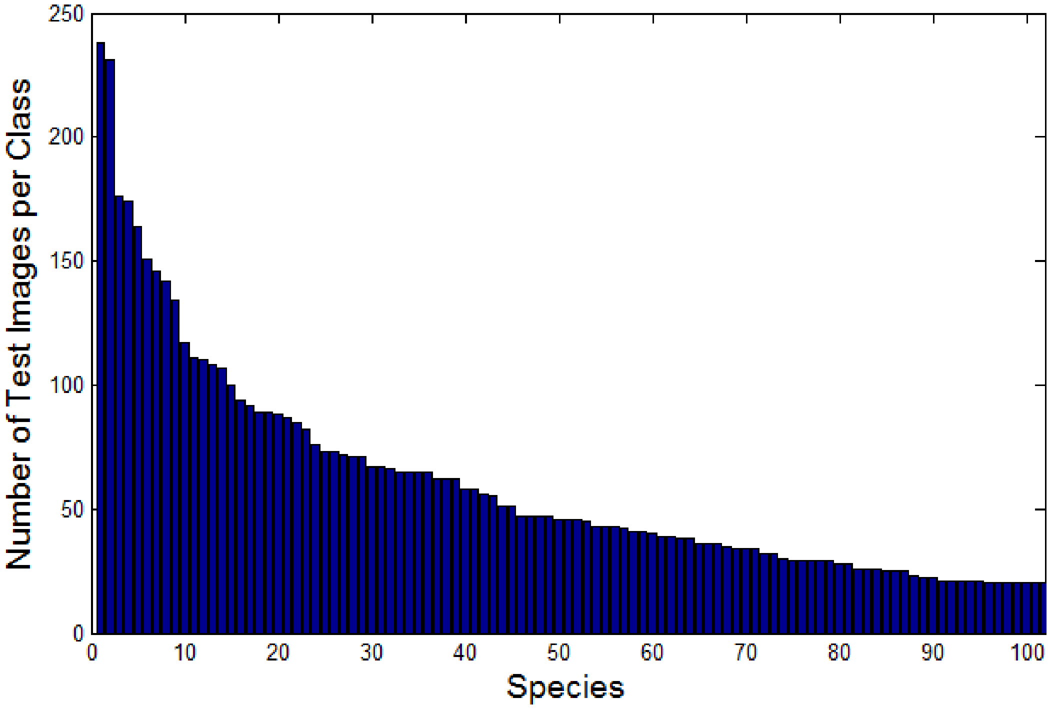
[Figure 3.6] In detail block diagram of our approach. (a) Block diagram: Dense SIFT descriptors are passed through the stacking block where descriptors from each plane are stacked independently whereas MSD-SIFT and NO-HOG descriptors are passed through the cascading block to construct higher dimensional descriptors. (b) Detailed view of the "Cascading" block. (c) Detailed view of the "Stacking" block.

4. Experiment

In this section, we illustrate the comparative performance of our compound descriptors against conventional ones as well as recent literatures using multiple kernel learning on Oxford-102 dataset [3].

4.1 Dataset

Oxford 102 class flowers dataset is introduced by Nilsback and Zisserman [3]. It contains 102 species of flowers and a total of 8189 images. Each category consists of between 40 and 250 images. The species are chosen to be flowers commonly found in United Kingdom. Most of the images in this dataset were collected from the web whereas a few of them were acquired by taking pictures. The dataset is divided into a training set, and a test set. Training and validation sets contain 10 images from each class, in total 1020 images in both sets. Remaining 6149 images belong to the test set. The distribution of the test images over all the classes is shown in figure 4.



[Figure 4] Distribution of the number of test images over 102 classes.

4.2 Results

All the images are first cut according to the smallest bounding box enclosing the foreground segmentation and then are rescaled to the smallest dimension of 500 pixels before feature extraction [3, 4].

Table 4.1 shows the comparison of our composite descriptors against the conventional ones. There are significant improvements for all the descriptors; the highest one is about 21% for non-overlapping HOG and the lowest one is about 7% for HSV and 8% and 10% for multi-scale dense SIFT and dense SIFT, respectively. Hence, the composite structure of our descriptors may be preferred not only in flower classification, but also in

[Table 1] Comparison Between Conventional and Composite Descriptors

Features	Recognition Rate (%)	
	Conventional Features	Composite Features
HSV	43.36	50.81
NO-HOG	37.03	58.64
D-SIFT	57.88	67.86
MSD-SIFT	68.27	76.06

other similar applications over the traditional ones.

Table 4.2 lists the recognition performance for different features and their combinations and table 4.3 shows the comparison against the recent methods. We compare our method with those using the training set of Oxford dataset for their supervision to keep the same baseline. On this baseline, our method provides slightly better recognition rate over the state-of-the-art of Murray and Perronnin [19].

Parameters: The optimum number of words is 900 for HSV feature in Oxford dataset. For NO-HOG, it is 1500 and for both D-SIFT and MSD-SIFT, this number is 3000. The block size for averaging in HSV colorspace is 3x3. For local standard deviation, range and entropy, the neighborhood chosen are 3x3, 3x3, and 9x9, respectively. In case of HOG, 8x8 square cells are used with the range of gradient orientation from -180 to +180 degree. Regular grid spacing of 5 pixels is used in the calculation of both

[Table 2] Recognition Performance of Features on Oxford Dataset

Features	Recognition Rate (%)
HSV	50.81
NO-HOG	58.64
D-SIFT	67.86
MSD-SIFT	76.06
HSV + NO-HOG	67.75
NO-HOG + D-SIFT	79.41
D-SIFT + MSD-SIFT	81.83
HSV + D-SIFT	76.26
HSV + MSD-SIFT	80.01
NO-HOG + MSD-SIFT	77.56
HSV + NO-HOG + D-SIFT	81.57
NO-HOG + D-SIFT + MSD-SIFT	82.86
HSV + NO-HOG + MSD-SIFT	80.60
HSV + NO-HOG + D-SIFT + MSD-SIFT	85.02

D-SIFT and MSD-SIFT. The radius for MSD-SIFT used is equal to 6. Scales equal to 4, 6, 8, and 10 are used in MSD-SIFT calculation. VLFEAT library [25] is used to compute the variants of SIFTs and bag-of-features modeling.

[Table 3] Performance Comparison on Oxford-102 Dataset

Method	Recognition Rate (%)
Nilsback and Zisserman [3]	72.8
Ito and Kubota [6]	74.8
Nilsback and Zisserman [9]	76.3
Chai et al., BiCos Method [5]	79.4
Chai et al., BiCos-MT Method [5]	80.00
Angelova and Zhu [18]	80.66
Aich and Lee (Our Previous Method) [14]	**81.05**
Murray and Perronin [19]	84.60
This method	**85.02**

5. Conclusion

In this paper, we propose an improved and more robust set of compound descriptors designed on the basis of the orthodox ones. We show significant performance improvement by each of them over conventional counterparts in the domain of flower classification. The computational time-complexity of these composite versions is similar to that of the traditional ones in parallel setups. Albeit we use them only for flower classification, these forms can be preferred in similar applications over the standard descriptors. However, although the features of our approach simply excel the original ones individually, their joint performance with multiple kernel learning is not that satisfactory due to high level of overlapping among the correctly classified samples by each feature. Therefore, future work may include developing the set of descriptors with less such overlapping.

References

[1] Everingham, M., Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88(2) (2009) 303–338.

[2] Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4) (April 2006) 594–611.

[3] Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics and Image Processing. (Dec 2008).

[4] Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2006) 1447–1454.

[5] Chai, Y., Lempitsky, V., Zisserman, A.: Bicos: A bi-level co-segmentation method for image classification. In: IEEE International Conference on Computer Vision. (2011).

[6] Ito, S., Kubota, S.: Object Classification Using Heterogeneous Co-

occurrence Features. In: Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V. Springer Berlin Heidelberg, Berlin, Heidelberg (2010) 701–714.

[7] Tang, L., Chen, J., Ye, J.: On multiple kernel learning with multiple labels. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence (2009).

[8] Nilsback, M.E., Zisserman, A.: Delving into the whorl of flower segmentation. In: British Machine Vision Conference. Volume 1. (2007) 570–579.

[9] Nilsback, M.E.: An Automatic Visual Flora – Segmentation and Classification of Flowers Images. PhD thesis, University of Oxford (2009).

[10] Nilsback, M.E., Zisserman, A.: Delving deeper into the whorl of flower segmentation. Image and Vision Computing (2009).

[11] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2) (November 2004) 91–110.

[12] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1. (June 2005) 886–893 vol. 1.

[13] Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA(2001).

[14] Aich, S., Lee, C.: A general vocabulary based approach for fine-grained object recognition. In: Image and Video Technology - 7th Pacific-Rim Symposium, PSIVT 2015, Auckland, New Zealand, November 25-27, 2015, Revised Selected Papers. (2015) 572–581.

[15] Rother, C., Kolmogorov, V., Blake, A.: "grabcut": Interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23(3) (August 2004) 309–314.

[16] Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vision 59(2) (September 2004) 167–181.

[17] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. (June 2010) 3360–3367.

[18] Angelova, A., Zhu, S.: Efficient object detection and segmentation for fine-grained recognition. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. (June 2013) 811–818.

[19] Murray, N., Perronnin, F.: Generalized max pooling. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. (June 2014) 2473–2480.

[20] Rabinovich, A., Vedaldi, A., Belongie, S.: Does image segmentation improve object categorization? Technical Report CS2007-090 (2007).

[21] Gonzalez, R.C., Woods, R.E.: Digital Image Processing. 2nd edn. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2001).

[22] Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual

categorization with bags of keypoints. In: In Workshop on Statistical Learning in Computer Vision, ECCV. (2004) 1–22.

[23] Mikolajczyk, K.: Detection of local features invariant to affine transfomations. PhD thesis, Institut National Polytechnique deGrenoble, France (2002).

[24] Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Proceedings of the 9th European Conference on Computer Vision - Volume Part IV. ECCV'06, Berlin, Heidelberg, Springer-Verlag (2006) 490–503.

[25] Vedaldi, A., Fulkerson, B.: Vlfeat: An open and portable library of computer vision algorithms. In: Proceedings of the 18th ACM International Conference on Multimedia. MM '10, New York, NY, USA, ACM (2010) 1469–1472.

[26] Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: IEEE International Conference on Computer Vision. (2007).

[27] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and

hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(5) (May 2011) 898–916.

[28] Sonnenburg, S., R¨atsch, G., Sch¨afer, C., Sch¨olkopf, B.: Large scale multiple kernel learning. J. Mach. Learn. Res. 7 (December 2006) 1531–1565.

[29] http://dovgalecs.com/blog/bag-of-visual-words-efficient-window-histogram-computation/

[30] Edelman, S., Intrator, N. and Poggio, T.: Complex cells and object recognition. Unpublished manuscript:
http://kybele.psych.cornell.edu/~edelman/archive.html

[31] http://www.scholarpedia.org/article/Receptive_field#Sherrington1906

[32] http://aishack.in/tutorials/sift-scale-invariant-feature-transform-features/

[33] Krizhevsky, A., Sutskever, I. and Hinton, G : Imagenet classification with deep convolutional neural networks. In: Proceedings of advances in Neural Information Processing Systems (NIPS), pp. 1106-1114.

[34] Razavian, A. S., Azizpour, H., Sullivan, J. and Carlson, S. : CNN features off-the-shelf: an astounding baseline for recognition. In Computing Research Repository (CoRR), 2014.

[35] Yoo, D., Park, S., Lee, J-Y. and Kweon, I-S. : Fisher kernel for deep neural activations. In Computing Research Repository (CoRR), 2014.

[36] Cortes, C. and Vapnik, V. : Support-vector networks. Mach Learn 20(3): 273-297.

[37] Bay, H., Tuytelaars, T. and Gool, L. J. V. : SURF: Speeded up robust features. In: Computer Vision – ECCV 2006: 9th European Conference on Computer Vision, Heraklion, Graz, Austria, May 7-13, 2006, Proceedings, Part I. Springer Berlin Heidelberg, Berlin, Heidelberg (2006) 404–417.

섬세한 어휘의 설명자 사용하여 꽃 종의 인식

AICH SHUBHRA

전남대학교대학원 전자컴퓨터공학과

(지도교수 이칠우)

(국문초록)

본 논문에서는 세밀한 인식으로 인해 더욱 다양해진 꽃들의 범주를 분류 문제에 대해 해결하고자 한다. 꽃마다의 다른 특징을 찾고 분류하는 다양한 설명 방법 중 단어-가방 모델들의 성능에 대해 조사한다. 이를 위해, 개별적으로 각 개체에 색상 평면을 가진 RGB 꽃 이미지에서 추가 정보를 도출 할 수 있는 더욱 정교한 방법인 로컬 이미지 통계들을 사용하는 설명방법을 제안한다. 위의 방법을 기반으로 대응점들을 보다 효율적으로 사용하여 더 넓은 범주의 꽃을 분류할 수 있는 방법을 설명한다. 여러 커널 학습 알고리즘을 사용하는 옥스포드 102 클래스 데이터 세트에서 제안된 설명방법을 통해 시각적 단어-가방 모델을 평가한다. 복합적 설명방법의 기반인 데이터 사전은 옥스포드-102 데이터 세트에 85.02 %의 정확도를 보여줌으로써 관리를 위한 데이터의 트레이닝 세트를 사용하는 다른 방법들보다 향상했음을 확인한다.

Keywords: Flower classification, compound visual vocabulary, local statistics, SIFT, HOG, multiple kernel learning.