A Visual Vocabulary for Flower Classification

Maria-Elena Nilsback and Andrew Zisserman
Robotics Research Group, Department of Engineering Science
University of Oxford, United Kingdom
men,az@robots.ox.ac.uk

Abstract

We investigate to what extent 'bag of visual words' models can be used to distinguish categories which have significant visual similarity. To this end we develop and optimize a nearest neighbour classifier architecture, which is evaluated on a very challenging database of flower images. The flower categories are chosen to be indistinguishable on colour alone (for example), and have considerable variation in shape, scale, and viewpoint.

We demonstrate that by developing a visual vocabulary that explicitly represents the various aspects *(colour, shape, and texture) that distinguish one flower from another, we can overcome the ambiguities that exist between flower categories. The novelty lies in the vocabulary used for each aspect, and how these vocabularies are combined into a final classifier. The various stages of the classifier (vocabulary selection and combination) are each optimized on a validation set.*

Results are presented on a dataset of 1360 images consisting of 17 flower species. It is shown that excellent performance can be achieved, far surpassing standard baseline algorithms using (for example) colour cues alone.

1. Introduction

There has been much recent success in using 'bag of features' or 'bag of visual words' models for object and scene classification [1, 3, 5, 7, 14, 15, 17]. In such methods the spatial organization of the features is not represented, only their frequency or occurrence is significant. Previous work dealing with object classification has focused on cases where the different object categories in general have little visual similarities (e.g. Caltech, 101), and models have tended to use off-the-shelf features (such as affine-Harris [12] detectors with SIFT [10] descriptors).

In this paper we investigate whether a carefully honed visual vocabulary can support object classification for categories that have a significant visual similarity (whilst still maintaining significant within-class variation).

To this end we introduce a new dataset consisting of different flower species. Classifying flowers is a difficult task even for humans – certainly harder than discriminating a car from a bicycle from a human. As can be seen from the examples in figure 2, in typical flower images there are huge variations in viewpoint and scale, illumination, partial occlusions, multiple instances etc. The cluttered backgrounds also makes the problem difficult as we risk classifying background content rather than the flower itself. Perhaps the greatest challenge arises from the intra-class vs inter-class variability, i.e. there is a smaller variation between images of different classes than within a class itself, and yet subtle differences between flowers determine their classification. In figure 1, for example, two of the flowers belong to the same category. Which ones?

Botanists use keys [6], where a series of questions need to be answered in order to classify flowers. In most cases some of the question are related to internal structure that can only by made visible by disecting the flower. For a visual object classification problem this is not possible. It is possible however to narrow down the choices to a short list of plausible flowers. Consequently, in this work as well as using the standard classification performance measures, we also use a measure on whether the correct classification is achieved within the top n ranked hypotheses. Measures of this type are very suitable for page based retrieval systems where the goal is to return a correct classification on the first page, but not necessarily as first ranked.

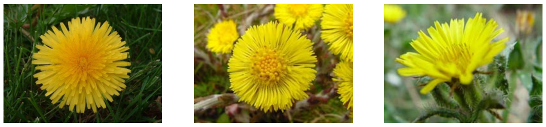


Figure 1. Three images from two different categories. The left and right images are both dandelions. The middle one is a colts' foot. The intra-class variation between the two images of dandelions is greater than the inter-class variation between the left dandelion and the colts' foot image.

What distinguishes one flower from another can sometimes be their shape, sometimes their colour and sometimes distinctive texture patterns. Mostly it is a combination of these three aspects. The challenge lies in finding a good representation for these aspects and a way of combining them that preserves the distinctiveness of each aspect, rather than averaging over them. However, flower species often have multiple values for an aspect. For example, despite their names, violets can be white as well as violet in colour, and 'blue bells' can be pink. This is quite exasperating, but indicates that any class representation will need to be 'multi-modal'.

1.1. Overview and performance measure

In the rest of this paper we develop a nearest neighbour classifier. The classifier involves a number of stages, starting with representing the three aspects as histograms of occurrences of visual words (a separate vocabulary is developed for each aspect) (section 2); then combining the histograms (section 3) into a single vocabulary. SIFT descriptors [10] on a regular grid are used to describe shape, HSV-values to describe colour, and MR-filters [16] to describe texture. Each is vector quantized to provide the visual words for that aspect. Each stage is separately optimized (in the manner of [4]).

Since we are mainly interested in being able to retrieve a short list of correct matches we optimize a performance cost to reflect this. Given a test image I_i^{test}, the classifier returns a ranked list of training images $I_j^{\text{train}}, j = 1, 2, ..., M$ with $j = 1$ being the highest ranked. Suppose the highest ranked correct classification is at $j = p$, then the performance score for I_i^{test} is

$$\begin{cases} w_p & \text{if } \ p \leq S \\ 0 & \text{otherwise} \end{cases}$$

where S is the length of the shortlist (here $S = 5$), and w_i is a weight which can be chosen to penalize lower ranks. If $w_i = 1 \ \forall \ i$ then the rank of the correctly classified image in the shortlist is irrelevant. We use a gentle fall off, of the form $w_i = 100 - 20\frac{i-1}{S-1}$, so that higher ranked images are rewarded slightly ($w_1 = 100, w_5 = 80$ for $S = 5$). Suppose the classifier is specified by a set of parameters $\boldsymbol{\theta}$, then the performance score over all test images is:

$$f(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} w_p^i & \text{if } \ p \leq S \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In essence, this is our utility/loss function, and we seek to maximize $f(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$. This optimization is carried out over a validation set in each of the following classification sections.

The performance of the developed classifier is compared to that of a baseline algorithm using colour histograms in section 4.

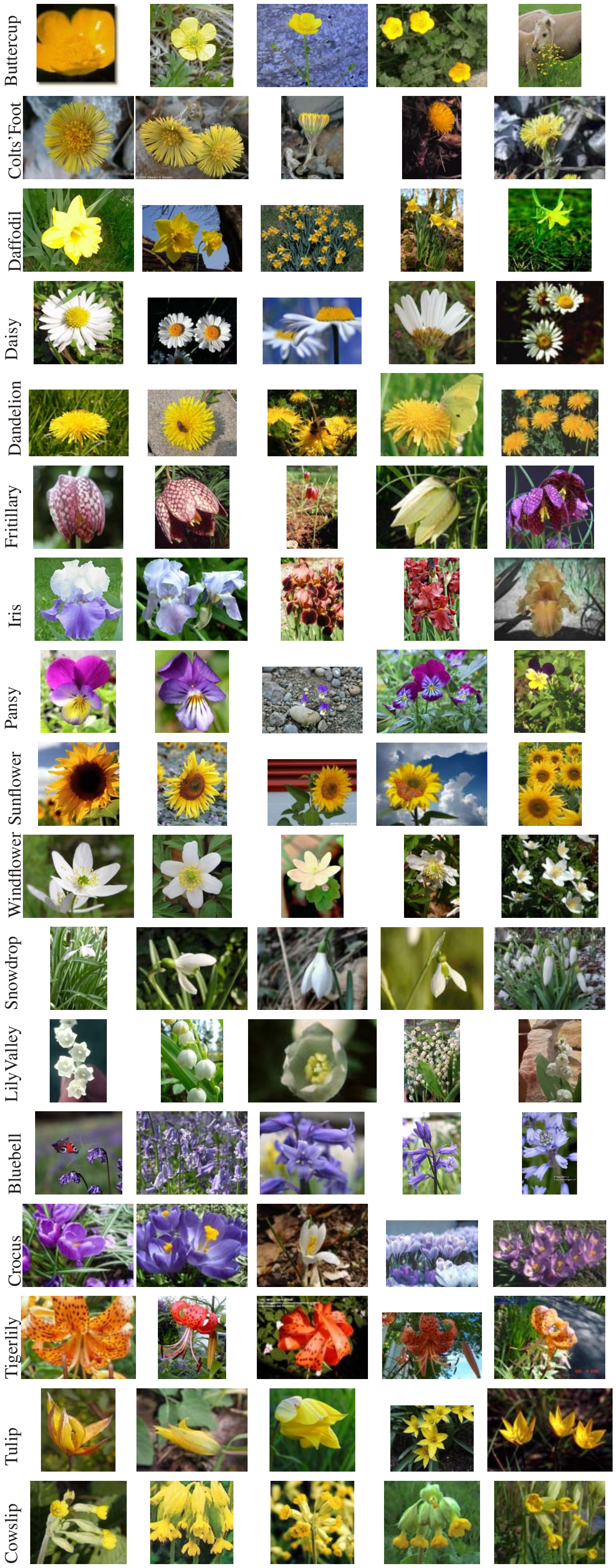


Figure 2. Images from the 17 category database. Each row shows 5 images from the same category. The first two columns in the top 10 rows show images from the restricted viewpoint set. Each category shows pose variation, scale changes, illumination variations, large intra-class variations and self-occlusion.

1.2. Datasets

The dataset introduced in this paper consists of 17 species of flowers with 80 images of each (figure 2). There are species that have a very unique visual appearance, for example fritillaries and tigerlilies, as well as species with very similar appearance, for example dandelions and coltsfeet. There are large viewpoint, scale, and illumination variations. The large intra-class variability and the sometimes small inter-class variability makes this dataset very challenging. The flower categories are deliberately chosen to have some ambiguity on each aspect. For example, some classes cannot be distinguished on colour alone (*e.g.* dandelion and buttercup), others cannot be distinguished on shape alone (*e.g.* daffodils and windflower). The flower images were retrieved from various websites, with some supplementary images from our own photographs.

Consistent viewpoint set: For the running example of the various stages of the classifier we do not use the full dataset, but instead consider only a subset. This consists of 10 species (figure 2) with 40 images of each. For each class the 40 images selected are somewhat easier than those of the full set, e.g. the flowers occupy more of the foreground or are orientated in a more consistent pose. We randomly select 3 splits into 20 training, 10 validation and 10 test images. The parameters are optimized on the validation set and tested on the test set. All images are resized so that the smallest dimension is 500 pixels.

Both the full and consistent viewpoint sets are available at http://www.robots.ox.ac.uk/∼vgg/data.html.

2. Creating a Flower Vocabulary

Like botanists we need to be able to answer certain questions in order to classify flowers correctly. The more similar the flowers, the more questions that need to be answered. The flowering parts of a flower can be either petals, tepals or sepals. For simplicity we will refer to these as petals. The petals give crucial information about the species of a flower. Some flowers have petals with very distinctive shape, some have very distinctive colour, some have very characteristic texture patterns, and some are characterized by a combination of these properties. We want to create a vocabulary that gives an accurate representation of each of these properties.

Flowers in images are often surrounded by greenery in the background. Hence, the background regions in images of two different flowers can be very similar. In order to avoid matching the green background region, rather than the desired foreground region, the image is segmented. The foreground and background RGB colour distributions are determined by labelling pixels in a subset of the training images as foreground (i.e. part of the flower), or background (i.e. part of the greenery). Given these foreground and background distributions, all images are automatically binary segmented using the contrast dependent prior MRF cost function of [2], optimized using graph cuts. Note, these distributions are common across all categories, rather than being particular to a species or image. This procedure produces clean segmentations in most cases. Figure 3 shows examples of segmented images. For the vocabulary optimization for colour and shape we compare the performance for both segmented and non-segmented images.

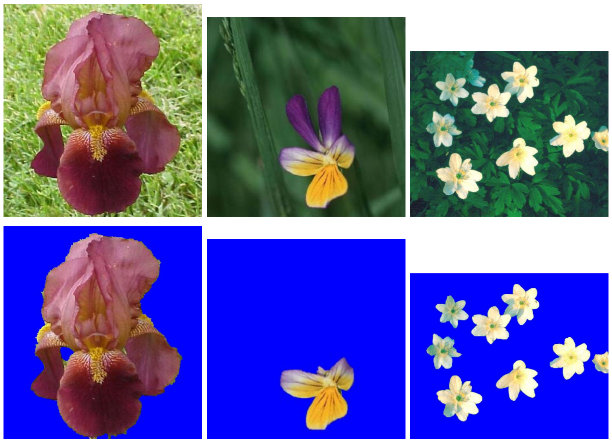


Figure 3. Segmented images. The top row shows the original images and the bottom the segmentation obtained. The flowers in the first and third column are almost perfectly segmented out from the background greenery. The middle column shows an example where part of the flower is missing from the segmentation – this problem occurs in less than 6% of the images.

2.1. Colour Vocabulary

We want to create a vocabulary to represent the colour of a flower. Some flowers exist in a wide variety of colours, but many have a distinctive colour. The colour of a flower can help narrow down the possible species, but it will not enable us to determine the exact species of the flower. For example, if a flower is yellow then it could be a daffodil or a dandelion, but it could not be a bluebell.

Images of flowers are often taken in natural outdoor scenes where the lighting varies with the weather and time of day. In addition, flowers are often more or less transparent, and specular highlights can make the flower appear lighter or even white. These environmental factors cause large variations in the measured colour, which in turn leads to confusion between classes.

One way to reduce the effect of illumination variations is to use a colour space which is less sensitive to it. Hence, we describe the colour using the HSV colour space. In order to obtain a good generalization, the HSV values for each pixel in the training images are clustered using k-means clustering. Given a set of cluster centres (visual words) w_i^c, $i = 1, 2, ..., V_c$, each image I_j, $j = 1, 2, ..., N$, is then

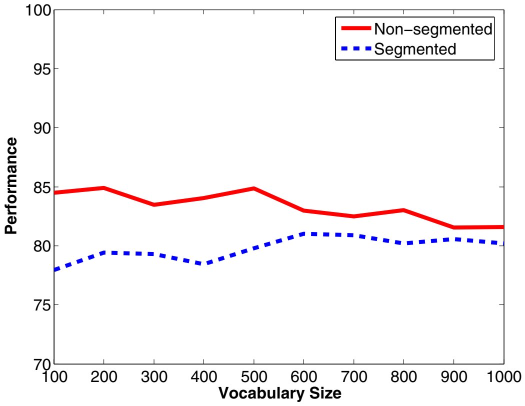


Figure 4. Performance (1) for the **colour** features. The results shown are averaged over three random permutation of training, validation and test sets. Best results are obtained with non-segmented images and 200 clusters, although the performance does not change much with the number of clusters, V_c.

represented by a V_c dimensional normalized frequency histogram $n(w^c|I_j)$. A novel test image is classified using a nearest neighbour classifier on the frequency histograms by determining c^*.

$$c^* = \arg\min_j d(n(w^c|I^{test}), n(w^c|I_j^{train})) \qquad (2)$$

where the distance $d(,)$ is computed using the χ^2 measure. We optimize the number of clusters, V_c, on the consistent viewpoint dataset. Figure 4 shows how the performance score (1) varies with the number of clusters. Results are presented for both segmented and non-segmented images. Perhaps surprisingly, the non-segmented images show better performance. This is because members of a flower species usually exist in similar habitats, thus making the background similar, and positively supporting the classification of the non-segmented images. However, in the full data set (as opposed to the rather restricted development set) this is not always the case and it is therefore better to segment the images. The best result using the segmented images is obtained with 500 clusters. The overall recognition rate is 55.3% for the first hypothesis and 84.3% for the fifth hypothesis (i.e. the flower is deemed correctly classified if one of the images in the top five retrieved has the correct classification).

2.2. Shape Vocabulary

The shape of individual petals, their configuration, and the overall shape of the flower can all be used to distinguish between flowers. In figure 5 it can be seen that although the overall shape of the windflower (left) and the buttercup (middle) are similar, the windflower's petals are more pointed. The daffodil (right) has petals more similar to that of the windflower, but the overall shape is very different due to the tubular shape corolla in the middle of the daffodil.

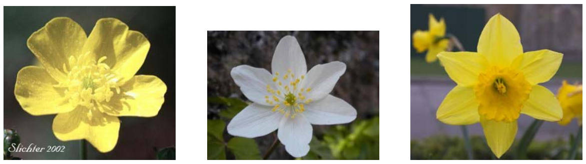


Figure 5. Images of similar shapes. Note that the windflower's (middle) petals are more pointy than the buttercup's (left). The daffodil (right) and the windflower have similar shaped petals, but are overall quite different due to the daffodil's tubular corolla.

Changes in viewpoint and occlusions of course change the perceived shape of the flower. The difficulty of describing the shape is increased by the natural deformations of a flower. The petals are often very soft and flexible and can bend, curl, twist etc., which makes the shape of a flower appear very different. The shape of a flower also changes with the age of the flower and petals might even fall off. For these reasons, the shape representation has been designed to be redundant – each petal is represented by multiple visual words for example, rather than representing each petal only once (attempting to count petals). This redundancy gives immunity to the occasional mis-classification, occluded or missing petal etc.

We want to describe the shape of each petal of a flower in the same way. Thus we need a rotation invariant descriptor. We compute SIFT descriptors [10] on a regular grid [4] and optimize over three parameters: the grid spacing M, with a range from 10 to 70 pixels; the support region for the SIFT computation with radius R ranging from 10 to 70 pixels; and finally, the number of clusters. We obtain $n(w^s|I)$ through vector quantization and classify the images in the same way as for the colour features.

Figure 6 shows how the performance score (1) changes on the development set when varying the size of the vocabulary, the radius and the step size. The best performance for the segmented images is obtained with 1000 words, a 25 pixel radius and a stepsize of 20 pixels. Note that the performance is highly dependent on the radius of the descriptor. The recognition rate for the first hypothesis is 82.7% and for the fifth hypothesis is 98.3%.

Figure 7 shows examples of some of the clusters obtained, and their spatial distribution. Note, that the shape-words are common across images and also within images. This intra-image grouping has some similarities to the Epitome representation of Jojic *et al.* [8], where an image is represented by a set of overlapping patches.

2.3. Texture Vocabulary

Some flowers have characteristic patterns on their petals. These patterns can be more distinctive, such as the pansy's stripes, the fritillary's checks or the tiger-lily's dots (figure

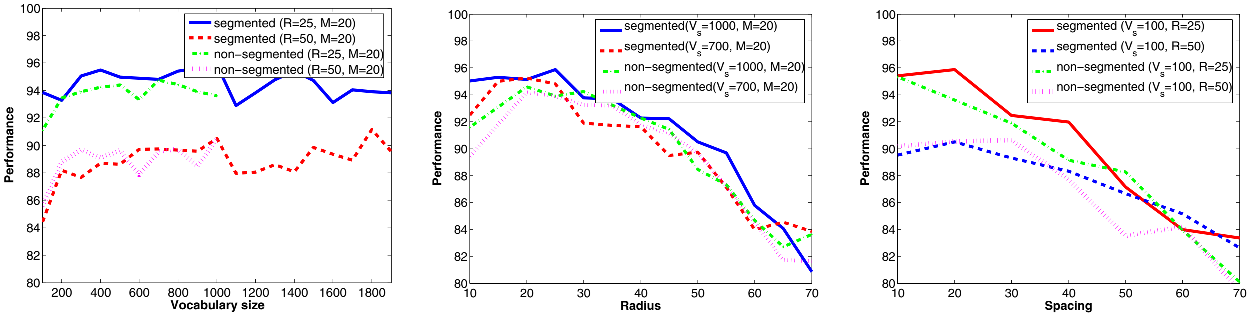


Figure 6. Performance (1) using **shape** vocabulary. Varying the number of clusters (V_s), the radius (in pixels) of the SIFT support window (R), and the spacing (in pixels) of the measurement grid (M).

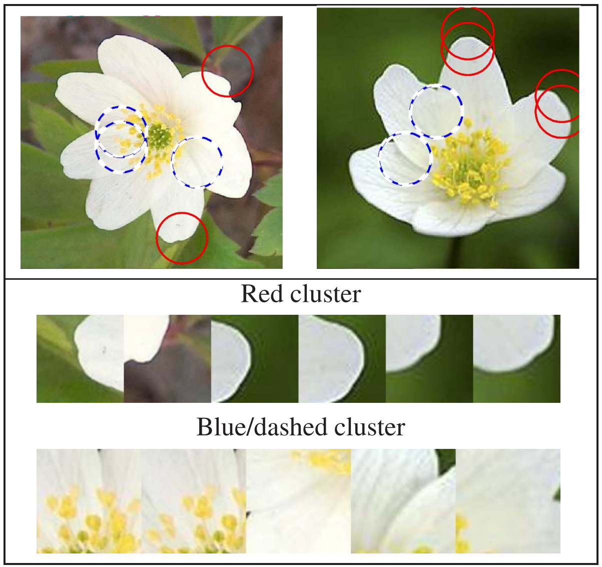


Figure 7. Two images from the Daffodil category and examples of regions described by the same words. All the circles of one colour correspond to the same word. The blue/dashed word represents petal intersections and the red word rounded petal ends. Note that the words detect similar petal parts in the same image (intra-image grouping) and also between flowers.

8), or more subtle in the form of characteristic veins in the petals. The subtle patterns are sometimes difficult to distinguish due to illumination conditions – a problem that also affects the appearance of more distinctive patterns.

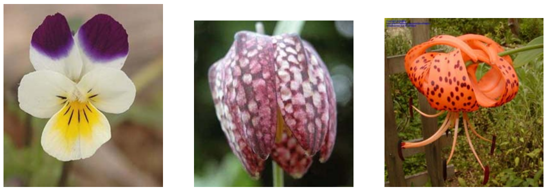


Figure 8. Flowers with distinctive patterns. From left to right: Pansy with distinctive stripes, fritillary with distinctive checks and tiger-lily with distinctive dots.

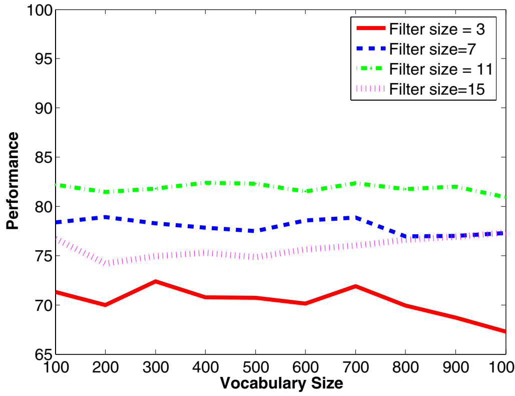


Figure 9. Performance (1) for the **texture** features on segmented images. Best results are obtained with filter size = 11 and 700 clusters.

We describe the texture by convolving the images with MR8 filter bank introduced by Varma and Zisserman [16]. The filter bank contains filter at multiple orientation. Rotation invariance is obtained by choosing the maximum response over orientations. We optimize over filters with square support regions of size $s = 3 - 19$ pixels. A vocabulary is created by clustering the descriptors and the frequency histograms $n(w^t|I)$ are obtained. Classification is done in the same way as with the colour features. Figure 4 shows how the performance varies with the number of clusters for different filter sizes. The best performance is obtained using 700 clusters and a filter of size 11. The recognition rate for the first hypothesis is 56.0% and for the fifth hypothesis it is 84.3%.

3. Combined Vocabulary

The discriminative power of each aspect varies for different species. Table 1 shows the confusion matrices for the different aspects for the consistent viewpoint set. Not surprisingly, it shows that some flowers are clearly distinguished by shape, *e.g.* daisies, some by colour, *e.g.* fritillaries and some by texture *e.g.* colts' feet, fritillaries. It also shows that some aspects are too similar for certain flowers, *e.g.* buttercups and daffodils get confused by colour, colts' feet and dandelions get confused by shape, and buttercups and irises get confused by texture. By combining the different aspects in a flexible manner one could expect to achieve improved performance. We combine the vocabularies for each aspect into a joint flower vocabulary, to obtain a joint frequency histogram $n(w|I)$. However, we have some freedom here because we do not need to give equal weight to each aspect – consider if one aspect had many more words than another, then on average the one with more words would dominate the distance in the nearest neighbour comparisons. We introduce a weight vector $\boldsymbol{\alpha}$, so that the combined histogram is:

$$n(w|I) = \begin{bmatrix} \alpha_s n(w^s|I) \\ \alpha_c n(w^c|I) \\ \alpha_t n(w^t|I) \end{bmatrix} \tag{3}$$

Since the final histogram is normalized there are only two independent parameters which represent two of the ratios in $\alpha_s : \alpha_c : \alpha_t$.

We learn the weights, $\boldsymbol{\alpha}$, on the consistent viewpoint set by maximizing the performance score of (1), here $f(\boldsymbol{\alpha})$, on the validation set. The performance is evaluated on the test set.

We start by combining the two aspects which are most useful for classifying the flowers, i.e. the shape and texture. Figure 10 shows $f(\boldsymbol{\alpha})$ for varying $\boldsymbol{\alpha}$'s. We keep $\alpha_s = 1$ fixed. Best performance is achieved with $\alpha_t = 0.8$. This means that the performance is best when texture has almost the same influence as shape. Combining shape and colour, however, leads to a superior performance. This is because the colour and shape complement each other better, whilst shape and texture often have the same confusions. The best performance for combining shape and colour is achieved when $\alpha_s = 1$ and $\alpha_c = 0.4$, i.e. when colour has less then half the influence of shape. The best performance is achieved by combining all aspects with $\alpha_s = 1.0$, $\alpha_c = 0.4$ and $\alpha_t = 1.0$. These results indicate that we have successfully combined the vocabularies – the joint performance exceeds the best performance of each of the separate vocabularies, i.e. we are not simply averaging over their separate classifications (which would deliver a performance somewhere between the best (shape) and worst (colour) aspect). Figure 11 shows an instance where both shape and colour misclassify an object but their combination classifies it correctly.

Discussion: The problem of combining classifications based on each aspect is similar to that of combining independent classifiers [11]. The $\boldsymbol{\alpha}$ weighting gives a linear combination of distance functions (as used in [11]). To see this, consider the form of the nearest neighbour classifier (2). Since in our case $d(\alpha\mathbf{x}, \alpha\mathbf{y}) = \alpha d(\mathbf{x}, \mathbf{y})$, (2) becomes

Colour										
buttercup	33.33	30.00	10.00	3.33	10.00		3.33		10.00	
colts foot	6.67	60.00			26.67		3.33		3.33	
daffodil	10.00	10.00	30.00		40.00				3.33	6.67
daisy				56.67			6.67			36.67
dandelion	3.33	40.00	13.33		30.00				13.33	
fritillary					3.33	93.33	3.33			
iris		6.67	6.67	13.33		13.33	36.67	10.00		13.33
pansy		3.33		13.33	3.33		23.33	46.67		10.00
sunflower	6.67	30.00			20.00				43.33	
windflower				30.00			10.00			60.00
Shape										
buttercup	70.00	3.33					13.33	13.33		
colts foot	3.33	63.33		10.00	23.33					
daffodil	3.33		60.00			13.33	16.67			6.67
daisy				96.67	3.33					
dandelion	3.33	16.67	3.33	3.33	73.33					
fritillary						90.00	3.33	6.67		
iris	16.67		6.67				70.00	3.33		3.33
pansy	16.67		6.67				16.67	53.33		6.67
sunflower				10.00		3.33	3.33		83.33	
windflower			3.33					3.33		93.33
Texture										
buttercup	46.67		10.00		10.00		23.33	10.00		
colts foot		86.67			13.33					
daffodil	6.67		60.00				6.67	6.67	13.33	6.67
daisy	6.67	10.00		50.00	20.00		3.33		10.00	
dandelion		33.33		3.33	46.67		3.33	6.67	3.33	3.33
fritillary		3.33			13.33	80.00	3.33			
iris	13.33	3.33	13.33		6.67		50.00	3.33	3.33	6.67
pansy	10.00		13.33	3.33		6.67	10.00	50.00	6.67	
sunflower	10.00	13.33	10.00	13.33	6.67		6.67		40.00	
windflower	13.33		6.67	10.00	6.67	3.33	3.33	6.67		50.00

Table 1. Confusion matrices for the first hypothesis of the different aspects on the consistent viewpoint dataset. The recognition rate is 75.3% for shape, 56.0% for texture and 49.0% for colour, compared to a chance rate of 10%.

$$\begin{aligned} c^* = \quad \arg\min_j \{ \quad & \alpha_s d(n(w^s|I^{test}), n(w^s|I_j^{train})) \\ & +\alpha_c d(n(w^c|I^{test}), n(w^c|I_j^{train})) \\ & +\alpha_t d(n(w^t|I^{test}), n(w^t|I_j^{train}))\} \end{aligned}$$

It is likely that learning weights for each class would increase the performance of the classification system, for example by learning a confusion matrix over all classes for each aspect. However, as the number of classes increases this becomes computationally intensive.

4. Results

In this section we present the results on the full 1360 image dataset consisting of 80 images for each of 17 cat-

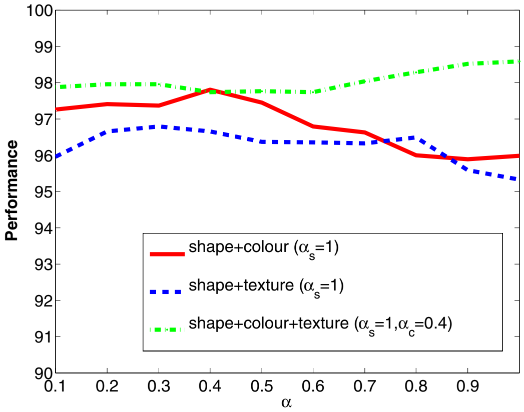


Figure 10. Performance for combining shape, colour and texture. The blue/dashed curve shows performance when varying α_t for a combination of shape and texture only, and the red/solid curve shows performance when varying α_c for a combination of shape and colour only. The best performance is obtained by combining all features (green/dot-dashed curve) with $\alpha_s = 1.0$, $\alpha_c = 0.4$ and $\alpha_t = 1.0$.

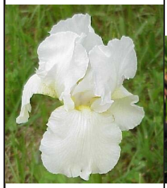
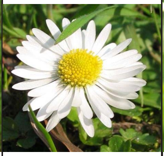
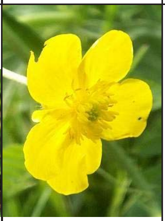
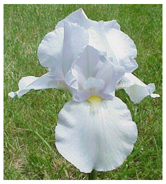
Test Image	1st Hyp Colour	1st Hyp Shape	1st Hyp Combined
			

Figure 11. Flowers misclassified by single aspect but correctly classified by their combination. From left to right: Original test image (Iris), first hypothesis for colour (Daisy), first hypothesis for shape (Buttercup) and first hypothesis combined (Iris).

egories. We use 40 training images, 20 validation images and 20 test images for each class. This dataset is substantially more difficult than the consistent viewpoint set. There are extreme scale differences and viewpoint variations, and also many missing and occluded parts. Since the datasets have significant differences we relearn the vocabularies in the same manner as for the consistent viewpoint set and optimize over the parameters.

Figure 12 shows the performance according to (1). The performance for the shape features are shown for R=25 and M=20. The colour features achieve a performance of 73.7% and the shape features achieve a performance of 71.8%, both with 800 clusters. Note that the colour features are performing better than shape on the larger set. This is probably because the development set has proportionally more instances of similar coloured flowers, and also because of the larger scale variations in the full set – which presents a challenge for the shape feature. The texture performs very poorly. This is because the proportion of classes distinguishable by texture is very small, and the texture features also suffer due to large scale variations. We achieve our best performance by combining shape and colour with $\alpha_s = 1$ and $\alpha_c = 1$. The performance according to (1) is 81.3%, a very respectable value for a dataset of this difficulty. Although the texture aspect has become redundant, the final classifier clearly demonstrates that a more robust system is achieved by combining aspects. Figure 13 shows a typical misclassification – illustrating the difficulty of the task – and figure 14 shows a few examples of correctly classified flowers.

We compare the classifier to a baseline algorithm using RGB colour histograms computed for each 10×10 pixel region in an image. Classification is again by nearest neighbours using the χ^2 distance measure for the histograms. The baseline performance is 55.7%, substantially below that of the combined aspect classifier.

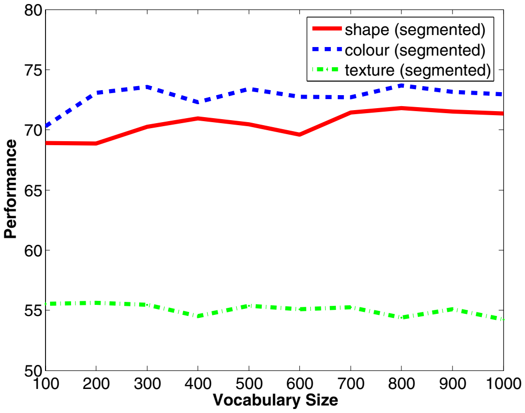


Figure 12. Performance for shape, colour and texture on the full datasets for different vocabulary sizes

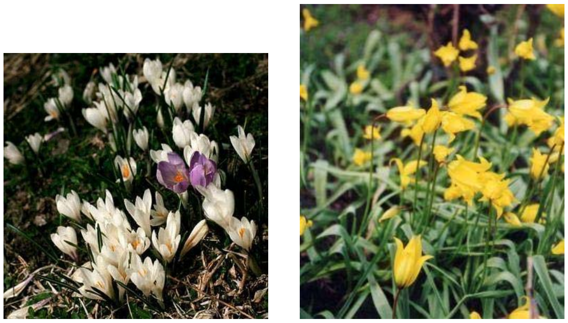


Figure 13. Misclassified image. A test image (left) of crocuses which is misclassified as wild tulips (right). This particular example also shows that there are images where it is difficult to distinguish the shape of the flower.

5. Discussion

We could have approached the problem of flower classification by building specially crafted descriptors for flow-

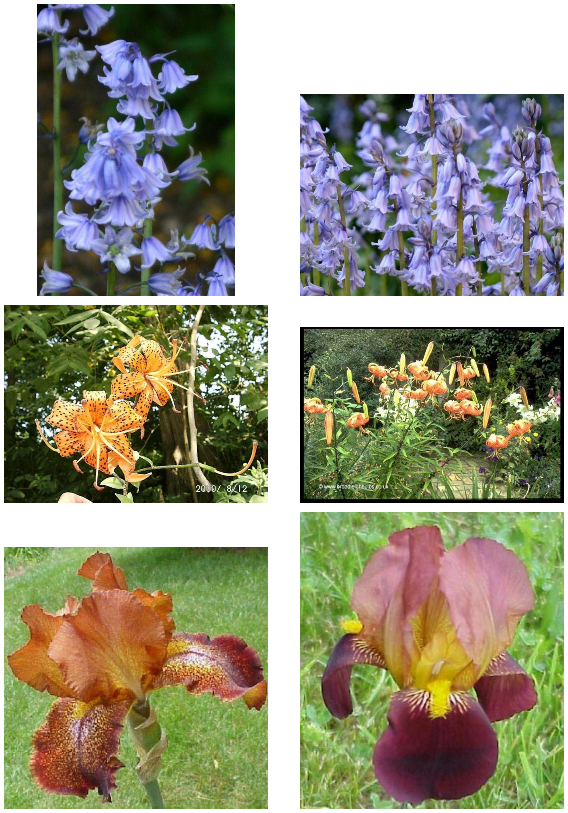


Figure 14. Examples of correctly classified images. The left column shows the test image and the right its closest match. Top: bluebells, middle: tigerlilies, and bottom: irises.

ers, for example a detector that could segment out petals, a stamen detector, an aster detector etc, with associated specialized descriptors. Indeed, such descriptors have already been developed for classifying based on scanned leaf shape [9, 13]. Instead of employing such explicit models, we have shown that more general purpose descriptors are sufficient – at least for a database with this level of difficulty. Tuning the vocabulary and combining vocabularies for several aspects results in a significant performance boost, with the final classifier having superior performance to each of the individual ones. The principal challenges now are coping with significant scale changes, and also coping with a varying number of instances – where a test image may contain a single flower or ten or more instances.

Acknowledgements

Thanks to the many friends who patiently gave flower identification training. Thanks to M. Pawan Kumar for providing the graph cut code. This work was funded by the the EC Visiontrain Marie-Curie network and EC project CLASS.

References

[1] A. Bosch, A. Zisserman, and J. Muñoz. Scene classification via plsa. In *Proc. ECCV*, 2006.

[2] Y. Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *Proc. ICCV*, volume 2, pages 105–112, 2001.

[3] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[4] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[5] G. Dorkó and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Proc. ICCV*, 2003.

[6] T. Elpel. *Botany in a Day*. HOPS Press, 2000.

[7] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, Jun 2005.

[8] N. Jojic, B. Frey, and A. Kannan. Epitomic analysis to appearance and shape. In *Proc. ICCV*, 2003.

[9] H. Ling and D. W. Jacobs. Using the inner-distance for classification of articulated shapes. In *Proc. CVPR*, volume 2, pages 719–726, 2005.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[11] S. Mahamud and M. Hebert. The optimal distance measure for object detection. In *Proc. CVPR*, volume 1, pages 248–258, 2003.

[12] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 1(60):63–86, 2004.

[13] F. Mokhtarian and S. Abbasi. Matching shapes with self-intersections: Application to leaf classification. *IEEE Transactions on Image Processing*, 13(5), 2004.

[14] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modelling scenes with local descriptors and latent aspects. In *Proc. ICCV*, 2005.

[15] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. In *Proc. ICCV*, 2005.

[16] M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In *Proc. ECCV*, volume 3, pages 255–271. Springer-Verlag, May 2002.

[17] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object catergories: An in-depth study. Technical Report, INRIA Rhône-Alpes, 2005.