Invariant Image-Based Species Classification of Butterflies and Reef Fish

Hafeez Anwar
hanwar@caa.tuwien.ac.at
Sebastian Zambanini
zamba@caa.tuwien.ac.at
Martin Kampel
martin.kampel@tuwien.ac.at

Computer Vision Lab
Institute of Computer Aided Automation
Vienna University of Technology
Vienna, Austria

Abstract

We propose a framework for species-based image classification of butterflies and reef fish. To support such image-based classification, we use an image representation which enriches the famous bag-of-visual words (BoVWs) model with spatial information. This image representation is developed by encoding the global geometric relationships of visual words in the 2D image plane in a scale- and rotation-invariant manner. In this way, invariance is achieved to the most common variations found in the images of these animals as they can be imaged at different image locations, exhibit various in-plane orientations and have various scales in the images. The images in our butterfly and reef fish datasets belong to 30 species of each animal. We achieve better classification rates on both the datasets than the ordinary BoVWs model while still being invariant to the mentioned image variations. Our proposed image-based classification framework for butterfly and reef fish species can be considered as a helpful tool for scientific research, conversation and education.

1 Introduction

In this paper, we deal with species classification of butterflies and reef fish from images. The patterns and colors found on these animals make them the marvelous art pieces of nature. The species are visually distinguished from one another based on these patterns due to which they serve as a useful cue for species-based classification. The aim of the work presented in this paper is to develop an image representation on top of these visual cues to support species-based image classification.

There are thousands of various species of butterflies and moths that are grouped into 126 families [13]. Due to this large number of species and families, butterfly classification becomes a highly complex task that requires expert level knowledge. However, efficiency is of importance both for humans and machines when it comes to the processing of such a substantial number of animal species. The colors and patterns of their wings play a key role in their visual classification and thus can be used by an image-based automatic framework to support the species classification of butterflies.

Pages 5.1-5.8
https://dx.doi.org/10.5244/C.29.MVAB.5

Figure 1: Image variations caused by various factors in the images of butterflies and reef fish

The oceans are home to hundreds of reef fish species [9]. The reefs serve as sources of their food and shelter from the predators. However, the ecosystem of these reefs is under constant threat from water pollution. Due to the effects of contaminants in the water, reef fish exhibit behavior changes [12] which, if monitored, will give useful information for their timely protection. Therefore, a framework designed for the image-based classification of reef fish will serve as a first step towards the design and development of an image-based fish behavior monitoring system. In addition to that, such a classification framework can serve in a number of other application areas such as education (e.g. in schools) and entertainment (e.g. aquariums).

As the images used for the described scenarios are taken in the wild, a number of image variations have to be considered as shown in Figure 1. Unlike images of other animals such as cows [11] or horses [3], rotation differences are common in the images of butterflies and fish. Both are imaged in cluttered backgrounds caused by objects found in their respective habitats. Other image variations due to the scale and position differences of the objects of interest are also common. Consequently, we utilize an image representation which is highly insensitive to such image variations in order to support the species-based image classification of butterflies and reef fish.

Our recent work [2] uses images of 15 species of butterflies and reef fish for the evaluation of the proposed method that improves the bag-of-visual words (BoVWs) image representation by encoding the visual words in an invariant manner. The global image representation is based on the angles produced by the triangulation performed among the positions of identical visual words. Since the angles of a triangle are invariant to changes in scale, orientation and position, the global image representation based on these angles is scale-, translation- and rotation-invariant as well and thus is most suitable for the task of species-based classification of butterflies and reef fish. We extend this previous work in the following

two major directions:

1. The local features must support the scale- and rotation-invariant global image representation at the local level as it is built on top of their geometric relationships. The paper contributes to this theme by extensively evaluating four methods of extracting rotation-invariant local features. Among these methods, three are scale-invariant while one concatenates local features that are extracted at multiple scales.

2. We also contribute to the existing datasets. For butterflies, we assemble and use the most diverse dataset to date consisting of 30 species. The UIUC dataset [8] has images of 7 species and the images in Leeds butterfly dataset [7] belong to 10 species. Our reef fish dataset is also the most diverse dataset that contains images of 30 species obtained from the Internet. FishCLEF [6] dataset has 10 species while Huang *et al*. [5] use images of 15 species. Fish4Knowledge[1] is undoubtedly the most diverse repository but its videos are obtained from cameras mounted in reef thus suffering from other problems such as irregular illumination and occlusions.

2 Datasets

We collected images of 30 butterfly species from the Internet by using their biological species names as well as their commonly used names e.g. *"Vanessa cardui"* is also called *"Painted Lady"*. The retrieved images were then manually examined to select the images where the butterfly of interest is actually depicted. The total number of images in the butterfly dataset is 2415 where the number of images per specie ranges from 30 to 100. The exemplar images of the butterfly species are shown in Figure 2 where each image is cropped to depict the butterfly of interest. However, images in the actual dataset have variations due to background clutter, orientation and scale differences.

The images of 30 reef fish species obtained from Internet were manually filtered to select the ones depicting the fish of interest. Here, we also used either the biological names or the commonly used names to collect the images. The fish dataset consists of 1600 images where the number of images per specie ranges from 30 to 100. The exemplar images of the fish species are shown in Figure 3.

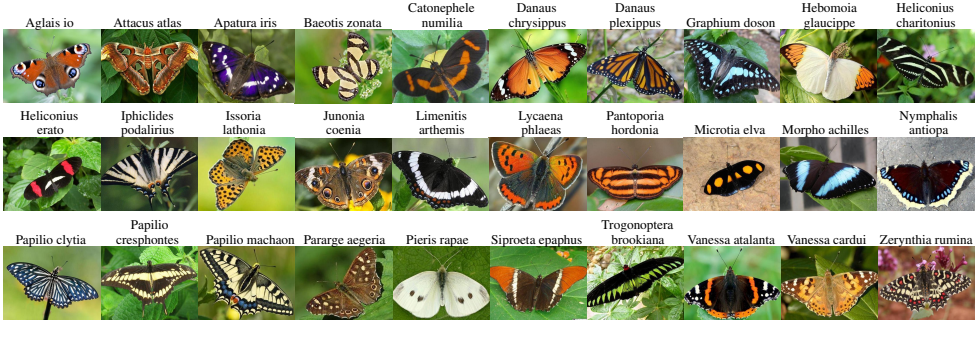

Figure 2: Exemplar images of the butterfly species

[1] http://groups.inf.ed.ac.uk/f4k/index.html

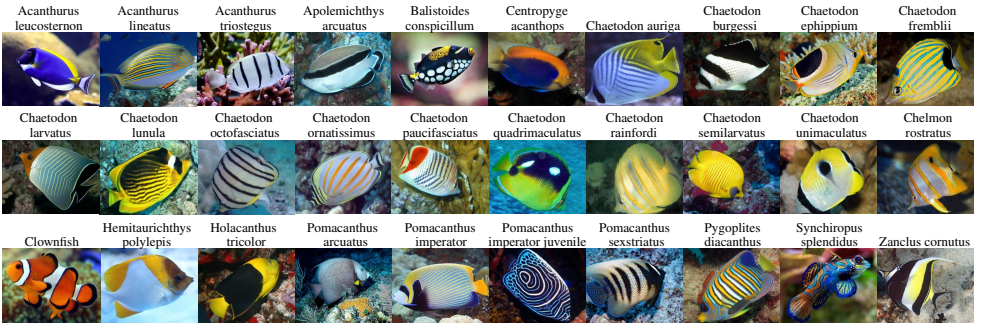

Figure 3: Exemplar images of the reef fish species

We mainly target the image variations caused by changes in object orientation, scale and translation and the severe background clutter found in the images of butterflies and reef fish. Due to this reason, we collected only those images from the Internet that suffer from these variations, unlike the images of Fish4Knowledge used by Huang *et al*. [5] where they are obtained from the cameras mounted in the reef thus suffering from other problems such as changes in illumination and occlusions.

3 Methodology

We propose to enrich the bag-of-visual words (BoVWs) model with spatial information in a manner that is invariant to changes in object scale and orientation. Khan *et al*. [7] proposed to use the angles made by pair-wise identical visual words (PIWs) to add spatial information to the BoVWs model. An image representation is then constructed on these angles by aggregating them in a pair-wise identical words angles histogram (PIWAH). We extended their idea to use the angles made by *triplets* of identical visual words (TIWs) [7]. The angles made by these triplets are then aggregated into the triplets of identical words angles histogram (TIWAH) to represent the images. The image representation based on this triangulation is invariant to changes in object scale, orientation and position. Since we use TIWAH to represent the images, a brief overview is given in the following.

In the BoVWs model, images are represented with a visual vocabulary which is made of visual words i.e. $voc = \{v_1, v_2, v_3, \ldots, v_M\}$ where M is the total number of words. A given image is represented as a set of local descriptors; $I = \{d_1, d_2, d_3, ..., d_N\}$ where N is the total number of descriptors. These descriptors are then mapped to the words of visual vocabulary using a similarity measure such as the Euclidean distance. The spatial position of a descriptor is given by its position on the dense sampling grid. Delaunay triangulation is used to triangulate the triplets belonging to a given word v_i. The angles from the triangulation are then aggregated in an angles histogram with bins between 0° and 180°. The i^{th} angles histogram $TIWAH_i$ calculated for the visual word v_i is used to replace the i^{th} bin of the histogram of visual words in such a way that the spatial information is added without losing the frequency information of v_i. Consequently, the given image is represented by combining $TIWAH_i$ of all the visual words. Since, this global encoding of the visual words is based on triangulation, it is scale-, translation- and rotation-invariant.

However, the local features must be discriminating enough to support the global image

representation. We propose to use SIFT [10] as a local rotation-invariant descriptor but to accommodate for local scale changes, we use the following extraction methods of SIFT.

- **Multi-scale SIFT:** The rotation-invariant SIFT features are densely extracted on a regular grid and at multiple pre-defined scales. These features are then concatenated.
- **Scale-less SIFT:** The SIFT descriptors extracted at multiple scales are combined to a single descriptor by Hassner *et al*. [9] which they call Scale-less SIFT (SLS). For each pixel, a set of SIFT descriptors is extracted at multiple scales. The Scale-less SIFT is then developed from these descriptors using subspace to point mapping techniques.
- **Difference of Gaussian (DoG) SIFT:** is based on the Difference-of-Gaussian interest point used by Lowe [10]. As a first step, interest points are detected using the Difference-of-Gaussian which is an approximation of the Laplacian-of-Gaussian. A non-maximal suppression is then performed to reject the low contrast points and those near edges. The remaining interest points are assigned orientations followed by the calculation of a 128-dimensional SIFT descriptor for each interest point.
- **Dense interest points (DIP) SIFT:** Tuytelaars [14] proposed a hybrid approach in which image patches are densely sampled on a regular grid and at multiple scales. Each patch is further refined both with respect to position and scale with some measure of interestingness such as the Laplacian. If a true local maximum is found within the patch limits, that point is considered as the center of the patch. If no maximum is found over the entire patch area, the center point of the patch is selected. A SIFT descriptor is then calculated for the patch centered on the local maximum.

4 Experiments

We perform experiments in a systematic manner. First we optimize for the number of scales in the Multi-scale SIFT on both the datasets using the ordinary BoVWs model. The Multi-scale SIFT with best performing number of scales is then evaluated along with the other three extraction methods on predefined vocabulary sizes. The best performing variant of SIFT is then used to compare the performances of PIWAH [7] and TIWAH. Each experiment is performed 10 times where in each run the datasets are randomly split into training and test sets. For classification, the histograms are precomputed with a Hellinger Kernel [16] and then used as feature vectors to a linear Support Vector Machine (SVM).

To reduce the effects of background clutter on the image representation, we manually generated segmentation masks for both the datasets and used them at the stage of vocabulary construction. This allows to use the local features from the foreground to construct the visual vocabulary, thus making it more discriminating and accurate as shown in [1].

To optimize for the number of scales on the given datasets, on a regular grid of pixel stride 10, rotation-invariant SIFT features are extracted and concatenated at 10 scales starting from a single scale of 2 and increasing the number of scales in such a way that any given scale is a $\sqrt{2}$ multiple of its predecessor. Results in Figure 4 demonstrate that in most cases 8 scales perform the best on both the datasets, while larger vocabulary sizes result in better classification rates. Therefore in the rest of the experiments, we use 8 scales in Multi-scale SIFT.

The four extraction methods of SIFT are evaluated on the current datasets. For SLS, we extract SIFT on a dense grid with pixel stride of 10 and from 20 scales linearly distributed

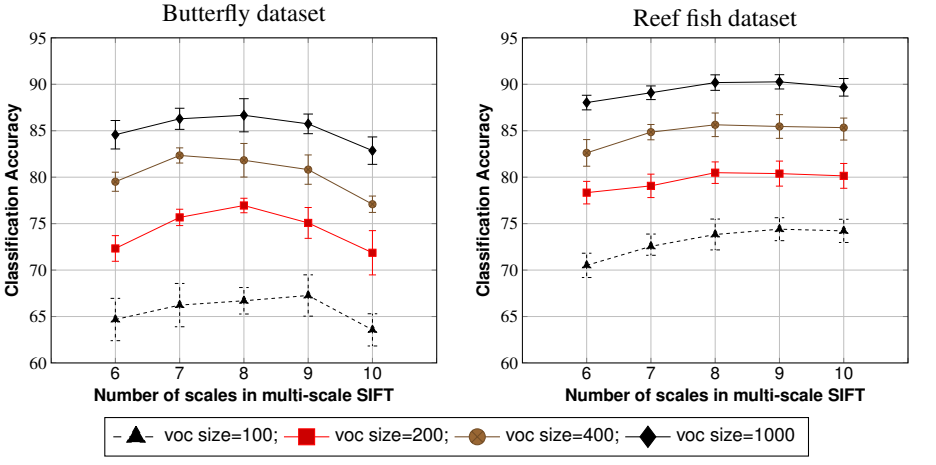

Figure 4: Results for the number of scales in Multi-scale SIFT on both the datasets for predefined vocabulary sizes. For clarity, the results of only 5 scales are shown

in the range [2, 32]. The limits of the range are based on the experimental results of Multi-scale SIFT. We use the default setting of `vl_sift` provided by the VLFEAT library [15] for DoG-SIFT. We also use the default settings for DIP-SIFT where the number of octaves is 4 with two scales per octave and pixel stride of the regular grid is 10. The results of experiments are shown in Figure 5 for the predefined vocabulary sizes. The performance of DoG-SIFT on the butterfly dataset is better than that on the reef fish dataset because it is based on a blob detector that localizes well on blobs found on the wings of butterflies. It performs worse on the reef fish dataset because most of the fish species have stripes instead of blobs. DIP-SIFT is also based on blob detector but accommodates for DoG-SIFT like

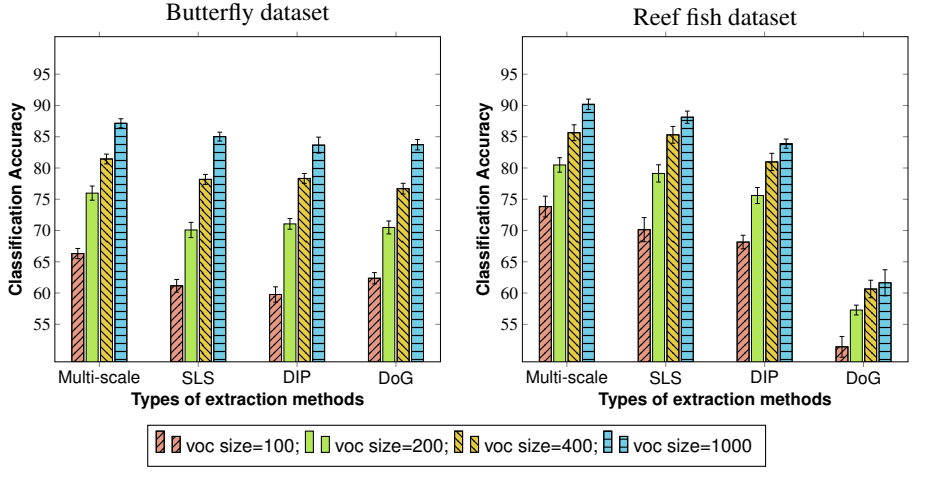

Figure 5: The classification rates achieved by extraction methods of SIFT on both the datasets for predefined vocabulary sizes

behavior due to its dense nature. However, it performs inferior to SLS and Multi-scale SIFT on both the datasets. The performance of SLS is near to that of the Multi-scale SIFT but it is not used in further experiments due to its expensive computations that involve extraction of SIFT from 20 scales and subspace to point mapping techniques.

Table 1 shows the performances of BoVWs, PIWAH and TIWAH for both the datasets on 4 vocabulary sizes. For the butterfly dataset, TIWAH outperforms BoVWs and PIWAH on smaller vocabulary sizes while at the vocabulary size of 1000 it performs marginally better. For the reef fish dataset it also performs better on smaller vocabulary sizes. Thus it can be concluded that the performance of TIWAH is more pronounced on smaller vocabularies on both the datasets. This phenomenon was also observed in the results of Khan *et al*. [7].

Table 1: Performances of BoVWs, PIWAH and TIWAH on both datasets for 4 vocabulary sizes

Datasets	Voc. Sizes	BoVWs		PIWAH		TIWAH	
		μ	σ	μ	σ	μ	σ
Butterfly	100	66.31%	2.32	69.64%	1.52	73.38%	1.54
	200	75.98%	1.61	75.56%	1.41	78.65%	2.08
	400	81.44%	1.77	81.03%	1.51	82.58%	1.53
	1000	87.10%	1.16	85.30%	1.31	87.15%	1.39
Reef fish	100	73.83%	1.14	75.60%	0.96	78.49%	0.95
	200	80.48%	1.58	81.66%	2.01	83.32%	1.77
	400	85.64%	1.09	85.69%	1.03	86.63%	0.87
	1000	90.18%	1.01	88.19%	1.4	89.54%	1.63

5 Conclusion

We presented an image-based framework for species classification of butterflies and reef fish on the most diverse datasets of both these animals. Due to the combined requirements of these two problems, the classification framework was supported by an image representation developed on top of the BoVWs model by enriching it with spatial information in a scale-, translation- and rotation-invariant manner. Furthermore, to support the global invariant image representation, four variants of SIFT are evaluated on the given datasets. The proposed framework outperformed the BoVWs model on both the datasets while still being robust to image rotations, translation and scale changes. However, it was observed to be more discriminating on smaller vocabulary sizes. In future, we plan to increase the size and diversity of our current datasets by adding images of more species of butterflies and fish. With such a huge and diverse dataset, we plan to investigate other sophisticated techniques for image classification such as deep learning.

References

[1] H. Anwar, S. Zambanini, and M. Kampel. Encoding spatial arrangements of visual words for rotation-invariant image classification. In *Proc. GCPR*, pages 407–416, 2014.

[2] H. Anwar, S. Zambanini, and M. Kampel. Efficient scale- and rotation-invariant encoding of visual words for image classification. *IEEE Signal Processing Letters*, 22(10): 1762 – 1765, 2015.

[3] E. Borenstein and S. Ullman. Learning to segment. In *Proc. ECCV*, pages 315–328, 2004.

[4] T. Hassner, V. Mayzels, and L. Zelnik-Manor. On sifts and their scales. In *Proc. CVPR*, 2012.

[5] P. X. Huang, B. J. Boom, and R. B. Fisher. GMM improves the reject option in hierarchical classification for fish recognition. In *Proc. WACV*, pages 371–376, 2014.

[6] A. Joly, H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W. Vellinga, R. Planque, A. Rauber, R. Fisher, and H. Müller. LifeCLEF 2014: multimedia life species identification challenges. In *Proc. CLEF 2014*, 2014.

[7] R. Khan, C. Barat, D. Muselet, and C. Ducottet. Spatial orientation of visual word pairs to improve bag-of-visual-words model. In *Proc. BMVC*, pages 1–11, 2012.

[8] S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *Proc. BMVC*, pages 779–788, 2004.

[9] E. Lieske and M. Robert. *Coral Reef Fishes: Indo-Pacific and Caribbean*. Princeton University Press, ISBN: 9780691089959, 2001.

[10] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, 2004.

[11] D Magee and R Boyle. Detecting lameness in livestock using re-sampling condensation and multi-stream cyclic hidden markov models. In *Proc. BMVC*, pages 34.1–34.10, 2000.

[12] S. D. Melvin and S. P. Wilson. The effects of environmental pollutants on complex fish behaviour: integrating behavioural and physiological indicators of toxicity. *Aquatic Toxicology*, 68(4):369 – 392, 2004.

[13] M. J. Scoble. *Geometrid Moths of the World: A Catalogue*. Apollo Books, ISBN: 8788757293, 1999.

[14] Tinne Tuytelaars. Dense interest points. In *Proc. CVPR*, pages 2281–2288, 2010.

[15] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[16] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE T-PAMI*, 34(3):480–492, 2012.

[17] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *Proc. BMVC*, pages 2.1–2.11, 2009.