Efficient Scale- and Rotation-Invariant Encoding of Visual Words for Image Classification

Hafeez Anwar, Sebastian Zambanini, and Martin Kampel

***Abstract*—The problem of incorporating spatial information to the bag-of-visual-words model for image classification is addressed in this letter. To incorporate such information, we propose to encode the global geometric relationships of the visual words in the $2D$ image plane in a scale- and rotation-invariant manner. This is established by measuring scale- and rotation-invariant geometrical properties given by triangles of identical visual words. Experimental results demonstrate that our proposed method is more robust to changes in scale and image rotations than the bag-of-visual words model on a butterfly and fish dataset.**

***Index Terms*—Image classification, object recognition, support vector machines.**

I. Introduction

THE bag of visual words model (BoVWs) has been used in a variety of problems such as scene classification [1], [2], large-scale image retrieval [3], [4] and object category recognition [5], [6]. In this technique, as a first step, local features such as SIFT [7] are collected from a set of images and quantized to form a vocabulary of the visual words. These visual words are then assigned to local features that are extracted from a given image based on a similarity measure. The image is then represented as a histogram of visual words. This histogram lacks the spatial information of the visual words which, if incorporated, results in better performance [1]. However, such spatial information must be robust to geometric transformations occurring in the image data, e.g. rotations [8].

The methods for the incorporation of spatial information to the BoVWs can broadly be divided into two groups. The methods in the first group split the image space into subspaces or tilings of various shapes and then from each tiling the statistics of visual words are collected. The most notable work in this group is the spatial pyramid matching (SPM) [1]. SPM divides the image space into hierarchically decreasing rectangular tilings. Weighted statistics of visual words from tilings at each level are then aggregated to achieve improved performance. Inspired by shape matching [9], log-polar tiling is used by Zhang *et al.* [10]. Single, multiple and multi-scale log-polar tilings are imposed on image space. From each sector of the log-polar tiling, statistics of visual words are extracted. They report improved performance over SPM on three benchmark datasets. Another recent method in this group also splits the image space into rectangular tilings which is called *word spatial arrangement*[11]. Defining the position of a given visual word as the origin, the image space is divided into four tilings. From each tiling, the information about the visual words is collected. This process is repeated for all the visual words of a given image. Finally, the information of all the four tilings is aggregated to represent the image. The methods in the second group encode the relationships among visual words. A notable work in this group is that of Khan *et al.* [12]. They propose to use the angles made by the positions of pairs of identical visual words with respect to the x-axis. A normalized histogram of angles is constructed to represent the image which they call pair-wise identical words angles histogram (PIWAH). However, all these methods are not robust to image rotations. To cope with image rotations in case of ancient coins, we proposed circular tiling [8], [13] which is only suitable for ancient coins as they can be automatically segmented [14] from the background. Circular tiling is not suitable in cases where automatic segmentation cannot be used.

The angles and ratios of the sides of a triangle are invariant to rotations, scale changes and translation. The triangular relationship of identical color patches is used by Tao and Grosky [15] to construct the so called *anglograms* for spatial color indexing and image retrieval. These anglograms are rotation, translation and scale invariant. To achieve triangular relationship among identical color patches, they use the *Delaunay triangulation* which is a well known and efficient triangulation [16] method of the computational geometry. To add spatial information to BoVWs, we build on both the works of Khan *et al.* [12] and Tao and Grosky [15]. We propose to encode geometric relationship of the identical visual words in an efficient scale- and rotation-invariant manner. We extend the idea of Khan *et al.* [12] to acquire the rotation-invariant geometric relationship among identical visual words. However, unlike them, we compute angles made by triplets of identical visual words instead of pairs of two identical words. From these angles, histograms are constructed in a similar manner as proposed by Khan *et al.* [12]. However, calculating angles for a huge number of unique triplets of identical visual words is a computationally expensive process. To reduce the calculation complexity, we use the Delaunay triangulation like Tao and Grosky [15]. To achieve rotation-invariance locally, we use dense SIFT features for which the dominant orientations are calculated. Besides rotation-invariant, Delaunay triangulation is also scale invariant as shown by Tao and Grosky [15]. In order to increase the discriminative power of the model on a local level, we extract dense SIFT features at several predefined scales. Following are

Manuscript received March 25, 2015; revised May 08, 2015; accepted May 11, 2015. Date of publication May 13, 2015; date of current version May 19, 2015. This work was supported by the Vienna PhD School of Informatics, Vienna University of Technology. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alexandre X. Falcao.

The authors are with the Computer Vision Lab, Institute of Computer Aided Automation, Vienna University of Technology, Vienna A-1040, Austria (e-mail: hafeez@caa.tuwien.ac.at; zamba@caa.tuwien.ac.at; martin.kampel@tuwien.ac.at).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2015.2432851

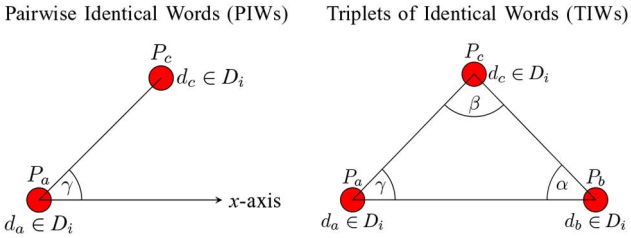

Fig. 1. PIWs and TIWs for the descriptors d_a, d_b and d_c at image positions P_a, P_b and P_c respectively. All descriptors belong to the same visual words D_i.

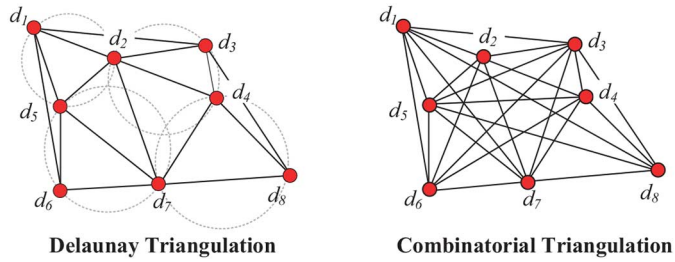

Fig. 2. Triangulation methods.

the major extensions of our recently published work [17] that we present in this paper.

1) The run-time required by the triangulation among the triplets of identical visual words is reduced by using the Delaunay triangulation.
2) Discriminating power of local features is increased by extracting them at several scales.
3) An extended dataset of butterflies is used.
4) A novel image dataset of 15 species of fish is used to evaluate the proposed method.

II. Scale- and Rotation-Invariant Histogram of Identical Visual Words

In the BoVWs model, similar image patches are assigned to the same visual word. Khan *et al.* [12] propose to use pairs of identical visual words (PIWs) for image description where a given pair consists of two identical words as shown in Fig. 1. The angles made by the positions of visual words with respect to the x-axis are calculated for all PIWs in a given image. These angles are then used to construct the PIW angle histogram (*PIWAH*) for image representation. Since the angles are computed with respect to the x-axis, their proposed method is not rotation-invariant. We modify their method to achieve rotation-invariant triangular relationship among identical visual words. We use three identical visual words in a given pair and denote it as Triplets of Identical Words (TIWs) as shown in Fig. 1. For image representation, the angles computed for each triplet are used to construct the TIW angles histogram (*TIWAH*).

In the BoVWs model, a visual vocabulary $voc=\{v_1, v_2, v_3, \ldots, v_M\}$ consists of M visual words. A given image is first represented as a set of descriptors; $I = \{d_1, d_2, d_3, \ldots, d_N\}$ where N is the total number of descriptors. A given descriptor d_k is then mapped to a visual word v_i using a similarity measures like the Euclidean distance,

$$v\left(d_k\right) = \arg\min_{v \in voc} \mathrm{Dist}\left(v, d_k\right) \tag{1}$$

where d_k is the kth descriptor in the image and $v(d_k)$ is the visual word assigned to this descriptor based on the distance $\mathrm{Dist}(v, d_k)$. The given image is then represented as the histogram of visual words where the number of bins of this histogram is equal M. Let D_i be the set of all descriptors mapped to a visual word v_i, then the ith bin of the histogram of visual words b_i, is the cardinality of the set D_i.

$$b_i = Card\left(D_i\right) \text{ where } D_i = \left\{d_k, k \in [1, \ldots, N] \,|\, v\left(d_k\right) = v_i\right\} \tag{2}$$

We previously proposed [17] to use all the distinct pairs of three descriptors from set D_i to calculate angles between the spatial positions of the descriptors as shown in Fig. 2. We call that method *combinatorial triangulation* as the triangulation is done for all the distinct triplets of descriptors belonging to a given visual word. The spatial position of a descriptor is given by its position on the dense sampling grid. The set of all TIWs related to a visual word v_i is defined as:

$$TIW_i = \left\{(P_a, P_b, P_c) \,|\, (d_a, d_b, d_c) \in D_i^3, d_a \neq d_b \neq d_c\right\} \tag{3}$$

where P_a, P_b and P_c are the spatial positions of the descriptors d_a, d_b and d_c respectively. The value of the ith bin of the histogram shows the frequency of the visual word v_i. Therefore, in case of *combinatorial triangulation*, the cardinality of TIW_i is the number of all possible triplets of distinct elements among the elements of D_i. The positions of the elements of each pair make a triangle. Calculating angles for such a huge number of triangles is time consuming. For instance, if the cardinality b_i of the set D_i is then the number of unique triplet combinations is 82160. Therefore, we propose to use the *Delaunay triangulation* where the number of triangles is much smaller. In *Delaunay triangulation*, the three points should not be collinear and the circumscribed circle defined by the three points should not contain any other point. The principles of the *Delaunay triangulation* significantly reduce the number of triangles for angle computation among TIWs. Fig. 2 shows both the *Delaunay* and the *combinatorial* triangulations. It can be observed that for 8 descriptors belonging to a visual word, *combinatorial triangulation* results in 56 triangles while the *Delaunay triangulation* results in 9 triangles. The angles histogram is built from the angles of Delaunay triangles with bins between 0° and 180°. The angles histogram for a specific word v_i is named as $TIWAH_i$. The ith bin of the histogram of visual words associated with visual word v_i is replaced with $TIWAH_i$ in such a way that the spatial information is added without losing the frequency information of v_i. Finally $TIWAH_i$ of all the visual words are combined to represent a given image.

$$TIWAH = (\psi_1 TIWAH_1, \psi_2 TIWAH_2, \ldots, \psi_M TIWAH_M)$$
$$\text{where } \psi_i = \frac{b_i}{||TIWAH_i||} \tag{4}$$

where ψ_i is the normalization coefficient. For a visual vocabulary of size M, if the number of bins in angles histogram is θ, then the size of the *TIWAH* is $M \cdot \theta$.

III. Experiments and Results

Experiments are performed on all the classes of the Leeds butterfly dataset [18] and five more classes which are 'Achilles Morpho', 'Common Jay', 'Machaon', 'Peacock' and 'Purple

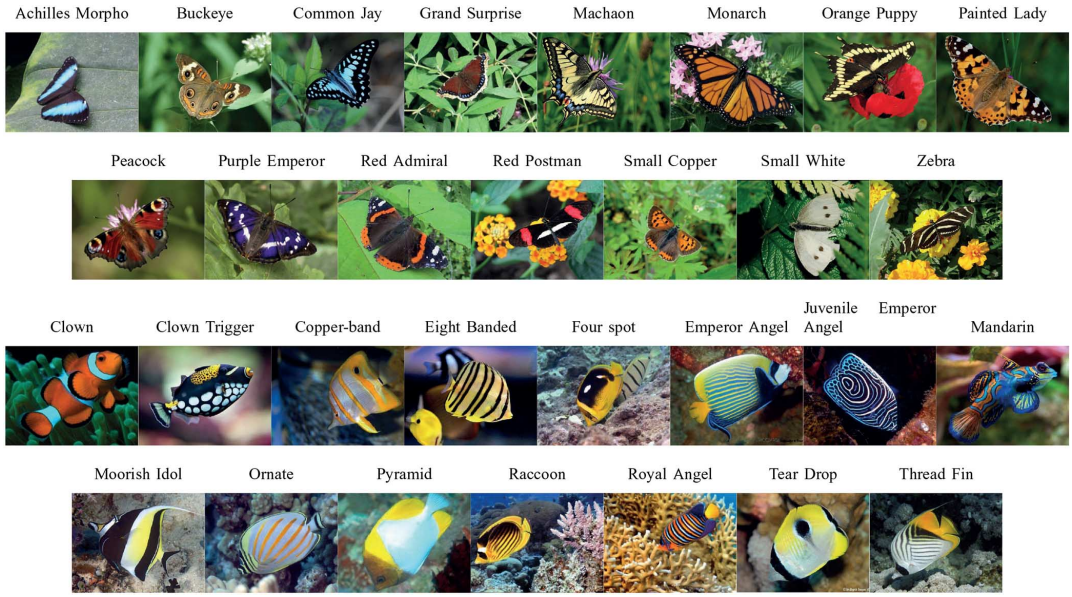

Fig. 3. Exemplar images of butterfly classes and Fish classes.

Emperor'. Furthermore, a novel dataset of butterfly fish species is created to evaluate the proposed method. In total, we use 15 classes in each dataset whose exemplar images are shown in Fig. 3. We use images of butterflies and fish to evaluate our proposed method because both of them can undergo changes in scale and in-plane rotations. In the butterflies dataset, the training set consists of 820 images while the test set consists of 351 images, while in the fish dataset 564 images are used for training and 242 for testing. For classification, we use the one-vs-all setting of SVM with Helinger kernel [19]. We extract SIFT features from images using a regular grid of pixel stride 10. To achieve rotation-invariance locally and enrich the features by multiple support regions, we compute the dominant orientation of each SIFT feature at several scales. We now give details of our experiments for various parameters.

A. Number of Scales for Local Features Extraction

On the given datasets, we optimize for the number of scales to extract the dense SIFT features. Starting from a single scale of 2, we extract features at 10 scales where a given scale is a $\sqrt{2}$ multiple of its predecessor. Rotation-invariant SIFT features are extracted and concatenated at predefined scales of $\{2, [2\ 4], [2\ 4\ 6], \cdots, [2\ 4\ 6\ 8\ 1\ 2\ 1\ 6\ 2\ 2\ 3\ 2\ 4\ 5\ 6\ 4]\}$. We previously showed that the use of segmentation masks at the stage of vocabulary construction enhances the discriminating nature of the vocabulary, thus resulting in a higher classification rate [17]. Here we also use segmentation masks to extract the foreground features for vocabulary construction. Results for the experiments on the number of scales for both the datasets are given in Fig. 4. The size of vocabulary is 200. Two main conclusions can be drawn from the results. First, for the butterflies dataset the maxima occur at 8 scales while for the fish dataset they occur at 7 scales. Second, on both the datasets $TIWAH$ clearly outperforms $PIWAH$ and the BoVWs model.

B. Run-Times and Classification Accuracies of the Triangulation Methods

To evaluate the efficiency of both the Delaunay and combinatorial triangulation schemes, we perform experiments to compare their classification accuracies and the computation time they take for a given set of images. The training and test sets are used to compare the classification accuracies of both the schemes. Experiments are performed 10 times and in each experiment the size of vocabulary is 200, the rotation-invariant local features are extracted at 8 scales for butterflies dataset and at 6 scales for fish dataset. The results averaged over the 10 classification runs are shown in Table I where it can observed that the Delaunay triangulation performs better than the Combinatorial triangulation. To find the time taken by each triangulation scheme for image representation, we select one image per class of butterflies and fish at random. Images are of standard size which is 640×480. In our experiments we also use a faster 'C' language implementation of the combinatorial triangulation. We denote this implementation as *combinatorialF* triangulation. We run the experiments 10 times on a single core and report the average time taken by the histogram representations using each triangulation method in Table I from which it can be noted that the Delaunay triangulation performs much efficiently than the combinatorial triangulation. To conclude, the Delaunay triangulation is more efficient than the combinatorial triangulation and also results in better classification rates on both the datasets.

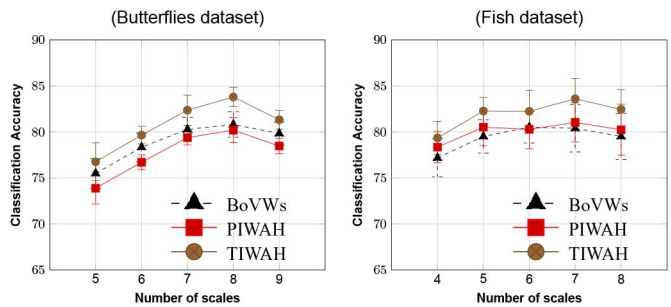

Fig. 4. Results for the number of scales on both the datasets. Due to the shortage of space, the results are shown for 5 to 9 scales. The mean performances are shown with their 95% confidence intervals.

TABLE I
CLASSIFICATION RATES AND TIME TAKEN IN SECONDS BY EACH TRIANGULATION METHOD

	Butterfly Dataset		Fish Dataset	
Methods	**Time**	**Accuracy**	**Time**	**Accuracy**
Combinatorial triangulation	1919.2	82.704	1858	81.68
CombinatorialF triangulation	39.24	–	39.7	–
Delaunay triangulation	**3.16**	**83.80**	**3.10**	**82.25**

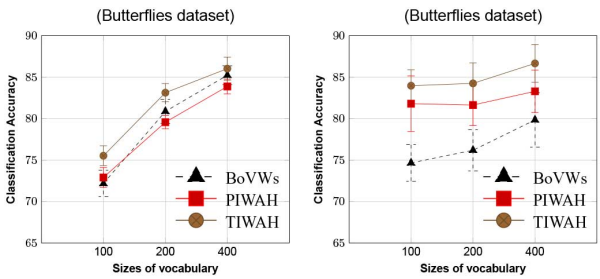

Fig. 5. Performances of BoVWs, PIWAH and TIWAH on predefined vocabulary sizes for both the datasets.

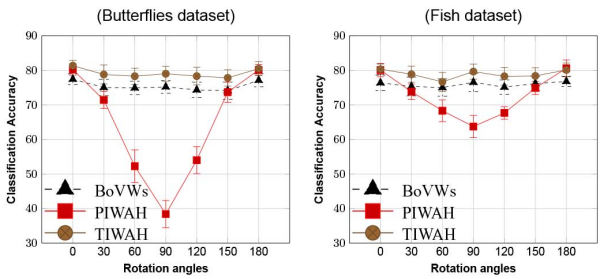

Fig. 6. Rotation-invariance evaluation of various methods on both the datasets.

C. Influence of Vocabulary Size

We also optimize for the sizes of vocabulary on our current datasets. The sizes of the vocabulary are predefined as $\{100, 200, 400\}$. The results are shown for all the methods in Fig. 5. Our proposed method outperforms the other two methods on both the datasets on all the vocabulary sizes.

D. Rotation-Invariance

Finally, we evaluate our proposed method for rotation-invariance along with BoVWs and *PIWAH*. All the images of the training set are roughly brought to the same orientation so that they have least rotation differences. To emphasize on rotation-invariance, the background is suppressed in all the images of the test set using the segmentation masks. All the images of the test set are rotated by predefined angles which are [30, 60, 90, 120, 150, 180]. Thus, our test set consists of 7 test subsets. For the butterflies dataset, rotation-invariant local features are extracted at 8 scales while for fish dataset they are extracted at 6 scales thus achieving rotation and scale invariance locally. Experiments are performed 10 times and the mean classification accuracy of each method is reported. A separate visual vocabulary is constructed at each iteration whose size is 200. Results shown in Fig. 6 indicate that *TIWAH* is insensitive to image rotations and outperforms the BoVWs model and the method proposed by Khan *et al.* [12] on both the datasets.

IV. CONCLUSION

An efficient method for spatial information incorporation to the commonly used bag of visual words model is proposed which is also invariant to changes in scale and image rotations. This is achieved via the scale- and rotation-invariant geometric relationships of the visual words in the 2D image space. Combinatorial explosion of the problem of encoding triangular relationships in a large set of visual words is prevented by using Delaunay triangulation in the selection process. Experimental results indicate that our proposed method not only outperforms the commonly used BoVWs model on two datasets but also efficiently achieves invariance to changes in scale and image rotations.

REFERENCES

[1] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 2169–2178.
[2] F.-F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 524–531.
[3] M. Perdoch, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval.," in *IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 9–16.
[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008.
[5] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," *ECCV*, pp. 1–22, 2004.
[6] J. Zhang, M. Marszaek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, 2007.
[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.
[8] H. Anwar, S. Zambanini, and M. Kampel, "Supporting ancient coin classification by image-based reverse side symbol recognition," in *Int. Conf. Computer Analysis on Images and Patterns (CAIP) (2)*, 2013, pp. 17–25.
[9] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 24, pp. 509–522, 2001.
[10] E. Zhang and M. Mayo, "Enhanced spatial pyramid matching using log-polar-based image subdivision and representation," in *Int. Conf. Digital Image Computing: Techniques and Applications (DICTA)*, 2010, pp. 208–213.
[11] O. A. B. Penatti, F. B. Silva, E. Valle, V. Gouet-Brunet, and R. da Silva Torres, "Visual word spatial arrangement for image retrieval and classification," *Patt. Recognit.*, vol. 47, no. 2, pp. 705–720, 2014.
[12] R. Khan, C. Barat, D. Muselet, and C. Ducottet, "Spatial orientation of visual word pairs to improve bag-of-visual-words model," in *Proc. Brit. Machine Vision Conf.*, 2012, pp. 1–11.
[13] H. Anwar, S. Zambanini, and M. Kampel, "Coarse-grained ancient coin classification using image-based reverse side motif recognition," *Mach. Vis. Applicat.*, vol. 26, no. 2–3, pp. 295–304, 2015.
[14] S. Zambanini and M. Kampel, "Robust automatic segmentation of ancient coins," in *Int. Conf. Computer Vision Theory and Applications (VISAPP)*, 2009, pp. 273–276.
[15] Y. Tao and W. I. Grosky, "Spatial color indexing using rotation, translation, and scale invariant anglograms," *Multimedia Tools Applicat.*, vol. 15, no. 3, pp. 247–268, 2001.
[16] S.-W. Cheng, T. K. Dey, and J. Shewchuk, *Delaunay Mesh Generation*, 1st ed. Boca Raton, FL, USA: Chapman & Hall/CRC, 2012.
[17] H. Anwar, S. Zambanini, and M. Kampel, "Encoding spatial arrangements of visual words for rotation-invariant image classification," in *German Conf. Pattern Recognition (GCPR)*, 2014, pp. 407–416.
[18] J. Wang, K. Markert, and M. Everingham, "Learning models for object recognition from natural language descriptions," in *Proc. Brit. Machine Vision Conf.*, 2009, pp. 2.1–2.11.
[19] A. Vedaldi and A. Zisserman, "Sparse kernel approximations for efficient classification and detection," in *IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012.