Ordered Minimum Distance Bag-of-Words Approach for Aerial Object Identification

Eren Unlu, Emmanuel Zenou
ISAE-SUPAERO
10 avenue Edouard Belin, BP 54032,
31055 Toulouse cedex 4, FRANCE
Eren.Unlu@isae-supaero.fr

Nicolas Riviere
ONERA
2 avenue Edouard Belin, BP 74025,
31055 Toulouse cedex 4, FRANCE

Abstract

Detecting potential aerial threats like drones with computer vision is at the paramount of interest for the protection of critical locations.This type of a system should prevent efficiently the false alarms caused by non-malign objects such as birds, which intrude the image plane. In this paper, we propose an improved version of a previously presented Speeded-up Robust Feature Transform (SURF) based algorithm, referred as Ordered Minimum Distance Bag-of-Words (omidBoW) to discriminate drones, birds and background from the patches, using an extended histogram set. We show that a SURF based object recognition can be well integrated to this context and this improved algorithm can increase accuracy up to 16% compared to regular bag-of-words approach.

1. Introduction

A remarkable advance has been witnessed on the Unmanned Aerial Vehicle (UAV) *-also known publicly as drone-* industry in the last decade, which has provided high accessibility to these devices by large number of regular customers [4]. Even though, this new technology trend has created wide range of possibilities from environmental protection to entertainment has also posed an alarming security fall. Due to their small sizes and low electromagnetic signatures, conventional security measures such as military radars are incapable of identifying these objects [5]. Therefore, a computer vision approach seems viable to automatically detect potential aerial threats such as drones. Several computer vision approaches using ordinary optical or acoustic cameras are proposed to detect potential UAV threats have been proposed in [5][9]. In this paper, we intend to use Speeded-up Robust Features (SURF) algorithm for discriminating image patches of birds, drones and background (such as clear sky, clouds, sky patch with non-uniform illumination). SURF is known for its robustness counter illumination variations and rotational invariance [3].

2. Related Work

SURF algorithm produces a vector of features (generally length of 64 floating point values), measuring intensity changes in different spatial directions, in the pre-defined periphery of each keypoint. Machine learning algorithms used in classification processes, such as Support Vector Machines (SVM), decision trees, Artificial Neural Networks (ANN), Linear Discriminant Analysis (LDA) etc., generally require a constant length input set. Hence, SURF algorithm is not usually considered for computer vision classification, as it may produce undetermined number of keypoints. In order to apply SURF algorithm to general object recognition task, authors in [10] develop a *Bag of Words* (BoW) approach. Their main idea is to use K-means clustering on all collected SURF feature vectors from training vectors to generate a *vocabulary* of *visual words*. At the end, they associate each detected SURF keypoint on a image to a visual word, by taking the minimum distance to a cluster centroid over K clusters. Authors in [10] defines $K = 500$ clusters (a dictionary of visual words) for general object recognition for a large dataset for high number of target classes. After creating visual word categories with K-means algorithm, they associate each SURF keypoint in each image with a word in the dictionary which it has the minimum distance to cluster centroid. For each image, after each keypoint is labeled by a word in the dictionary, a K length histogram is created, counting the percentage of each word in that particular image. This K length histogram is used for classifying images by applying machine learning algorithms.

3. Ordered Minimum Distance Bag-of-Words Approach

We develop a robust object classifier for a potential aerial target detector based on computer vision. Our system, which uses a regular optical camera, is assumed to firstly detect any moving object by using pixel-wise background subtraction techniques, such as Gaussian mixture modeled background. This approach is usually preferred for similar tasks [7][1]. Following this, a rectangular bounding box is drawn on the moving target, defining borders of it. The moving target bounding boxes can be tracked by specific algorithms such as Extended Kalman Filters etc. We present a new kind of BoW approach based on SURF features of moving target bounding box patches. We believe a SURF feature based classification constructed on a visual word vocabulary may be particularly useful in this aerial object recognition task. One reason for that the bounding boxes applied on a moving target (such as a drone or bird) contains high number of pixels of background. In this context, certain times background visuals can be detected as a moving target due to rapid illumination effects, moving clouds etc., causing false positives. If several keypoints corresponding to background, which is not of interest can be defined in the visual vocabulary, by evaluating histogram of words of a patch, it can be defined as a false positive background patch. In addition, using proven efficiency of SURF features aerial objects can be classified with high performance.

In [10], a general object classification is intended for a large dataset and high number of target classes. Authors have determined a vocabulary size of 500 words as optimal, a trade-off between over-fitting and inadequacy of representation. However, in our particular case of aerial target recognition, there are limited number of classes such as birds, drones, planes etc. We can also add background patches generated as positive false into this set. Therefore, we may except to have a much smaller vocabulary compared to general object recognition task. Hence, this algorithm can be applied more computationally efficient onto this case. Our contribution in this paper is an improved version of the SURF BoW algorithm, which uses an extended histogram of words. As mentioned previously, each keypoint is labeled to a word, which it has the minimum euclidean distance to the cluster centroid. Even though, this approach may be very effective in the general object recognition with a large number of visual words, a better method can be applied in specific limited object recognition tasks, such as aerial object recognition. In specialized recognition tasks, associating a keypoint just to the nearest cluster may cause loss of useful information. For instance, the farthest cluster can be very descriptive for a keypoint for a constrained context (e.g. choosing the least likely word for a keypoint).

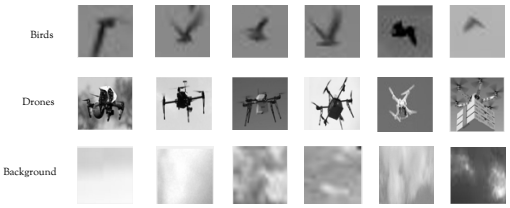

Figure 1. A few examples of the 64x64 grayscale bird, drone and background patches from the images we have collected from the internet.

Therefore, we have modified the BoW apporach by using extended set of histograms, which labels the words at each step choosing the ordered minimum distance. If there are K words in the vocabulary, our proposed feature vector is composed of K histograms (each length of K). As in the standard method, first histogram is the histogram of words, keypoints labeled with the nearest cluster. Second histogram is composed of keypoints labeled with the second nearest cluster (secondarily likely word) and ith histogram composed of the keypoints labeled with the ith nearest cluster. Therefore, by generating a new feature space based on BoW principles, we aim to capture visual characteristics of a keypoint as much as possible. Note that, this proposal is less efficient computationally (in the order of K^2), compared to the standard one which in the order of K. For large vocabularies the computational complexity of the proposed approach may be prohibitive, however this extension can be compromised for its informational augmentation in cases with smaller vocabularies.

In this work, we try to discriminate small image patches as bird, drone or background, using this new extended SURF omidBoW approach. We have collected different images of birds and drones in the sky from the internet. We have tried to have images of birds and drones which is small compared (few dozens of pixels) to background. Then, we have extracted these low resolution region of interest from the images. The reason for this is to imitate a real aerial object detection system as much as possible. These systems have a wide angle view and they should be able to detect objects at large distance, thus resulting in low resolution patches. However, our dataset also contains few number of higher resolution regions of interest. Before applying SURF algorithm into these patches, we have rescaled them to 64x64 pixels. Fig. 1 shows a few samples of bird, drone and background images. Especially, for background images, we have tried to collect patches with different scales as much as possible, in order to be able to characterize them in different bounding box sizes and shapes, which may be generated during the system operation.

4. Kernel Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), *also referred as Fisher's Discriminant Analysis (FDA)*, is a prominent classification algorithm [6]. An optimal weighted sum of the features of each individual is created to represent them as discriminative as possible, in terms of classes. These resulting representative values are called [2][11]. It is important to recall that, number of discriminative axes is equal to K-1, where K is the number of classes. Successful kernel versions can be developed of this algorithm [8]. 3 of the most popular kernels used in discriminant analysis are linear kernel, polynomial kernel (the linear data kernel elements are taken a power of a scalar) and a radial basis function kernel (Euclidian distance of individuals are measured and fed to a Gaussion function).

In this paper, we have evaluated linear kernel and 2nd degree polynomial kernel discriminant analysis, which we have acquired the best results. Normally, when the data is taken to the kernel space, it is not possible to measure the relative weights of the parameters on the result (how parameter contributes to the new scores *-as scores are weighted linear sum of the parameters*). This is because, in kernel space, we measure the combined similarity between individuals, where we lose the information on parameters. However, if linear kernel is used, with proper algebraic manipulation, it is possible to retrieve the weights of the parameters by multiplying data matrix with the resulting scores. As mentioned previously, with kernel linear discriminant analysis, we can have K-1 dimensions of scores, if there are K classes. Therefore, in our case with 3 classes, we have 2 dimensions of scores. In other words, each individual image is represented by 2 different scalars. And we have 2 different weight functions, which measure the relative contribution of each parameter on first and second scores. Another important point on this linear discriminant analysis procedure is the *regularization*. Before calculating the eigendecomposition, a proper regularization parameter is chosen to expand the training space, which ensures that overfitting does not occur. In our experiments, we have compared the regular BoW approach to our omidBoW approach and set two different optimal regularization parameter with k-fold analysis. However, it is important to state that this procedure does not finalize the classification and it requires an additional classifier such as SVM or simpler mathematical operations (fitting a boundry curve between classes) in order to label individuals on the generated discriminative plane. In this particular study, we have chosen to fit optimal boundry curves (2nd degree polynomials) between classes for a more intuitive representation.

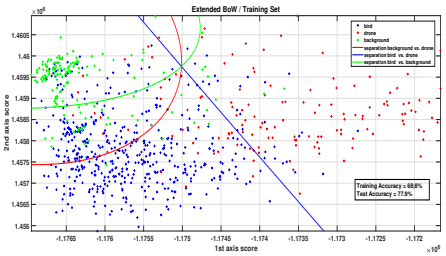

Figure 2. 1st and 2nd discriminant axes scores of training dataset samples and the quadratic boundry lines between classes, for our proposed extended omidBoW approach.

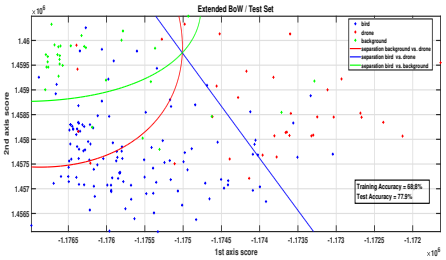

Figure 3. 1st and 2nd discriminant axes scores of test dataset samples and the quadratic boundry lines between classes, for our proposed extended omidBoW approach.

5. Experimental Evaluation

As mentioned previously, our context is much more limited compared to general object classification task, thus a much smaller dictionary length can be expected. Based on our experimentation, We have determined that a dictionary of 15 visual words is ideal. We have 654 bird images, 175 drones images and 1850 background images in our dataset. 80% of the patches from each class are used as training and 20% is used for testing. A 5-fold analysis is adopted. In total we have collected 2452 SURF keypoints among bird images, 1176 SURF keypoints among drone images, and 385 keypoints among background images. Note that, the number of detected SURF keypoints in background patches are remarkably low due to intensity uniformity. We have tested the performance of our omidBoW algorithm to regular BoW approach. The optimal regularization parameters are found for two algorithms, separately. The regularization parameter is chosen to maximize the overall testing dataset accuracy.

In order to emphasize the discrimination, we have taken negative log values of the histogram features. 0 values are kept as 0 in this process. As mentioned previously linear kernel and 2nd degree polynomial kernels are used. Fig. 2 and Fig. 3 show the resulting discriminant axes scores for our omidBoW algorithm, using linear kernel function for training and test sets, respectively. With our omidBoW ap-

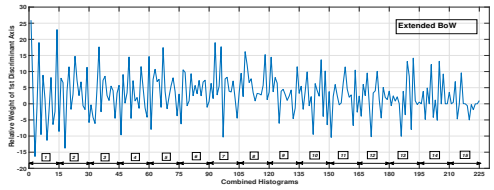

Figure 4. Weights of the parameters of linear kernel function with the extended omidBoW approach for the 1st discriminant axis.

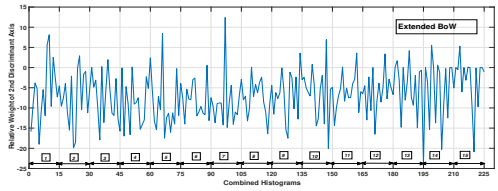

Figure 5. Weights of the parameters of linear kernel function with the extended omidBoW BoW approach for the 2nd discriminant axis.

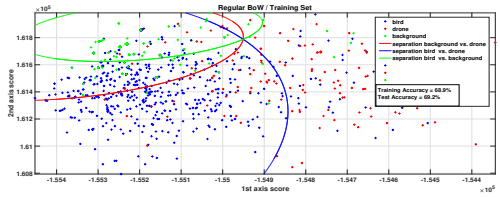

Figure 6. 1st and 2nd discriminant axes scores of training dataset samples and the quadratic boundry lines between classes, for regular BoW approach.

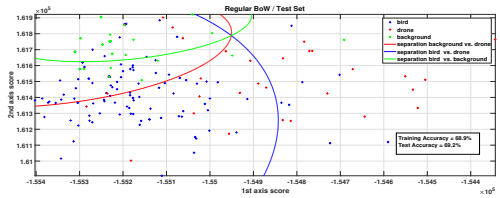

Figure 7. 1st and 2nd discriminant axes scores of test dataset samples and the quadratic boundry lines between classes, for regular BoW approach.

proach, we have achieved a 77.9% accuracy on testing set, with linear kernel function. As you can see from the figures, 1st axis mostly discriminates between birds and drones. As mentioned previously, for the linear kernel, we can manipulate scores to obtain relative weights of parameters for 1st and 2nd discriminant axes. Fig. 4 and Fig. 5 show the weights of 225 parameters (15 histograms, each with 15 visual words), respectively.

If we interpret these figures, we can see that generally words from each histogram of 15, has contributed to discrimination. This is a remarkable observation, as not only labeling the words with minimum distances to clusters, but labeling with them based on every possible distance configuration augments discriminative information significantly. For instance, from Fig. 4, we see that 1st and 2nd words' histogram values, based on the minimum distance labeling (1st histogram) contributes mostly to the discrimination on 1st axis (Note that, one of them discriminates in the negative direction). And we see that, 3rd word's histogram value, based on the 3rd minimum distance labeling (3rd histogram) contributes also significantly. Interestingly, on Fig.5, we observe that the histogram values of the words on the maximum distance labeling (15th histogram) contributes to the distinction in 2nd axis orientation in the negative direction. This is very important, as it can be interpreted as the effect of choosing the words, which are least likely (maximum distance) is a significant contributer. In a sense, we can state that negative direction on 2nd axis means, choosing the values from the 15th histogram mostly discriminates bird and drones counter the background patches.

Next, we present the results for the regular BoW approach in the literature, under same configuration for training and test datasets, in Fig. 6 and Fig. 7, respectively. We have achieved a 69.2% overall accuracy on test datasets, which signifies a 8.9% improvement with our proposed omidBoW algorithm. We have also tested our algorithm with Quadratic Kernel Function, where 2nd degree power is taken for kernel elements. Fig. 8 and Fig. 9 shows the 1st and 2nd discriminant axis scores based on our extended omidBoW algorithm, with a quadratic kernel. It can be seen that scores are spread in a different pattern compared to linear kernel case, due to higher dimensionality. We have achieved a 78.21% overall accuracy on test datasets. Fig. 10 and Fig. 11 shows the 1st and 2nd discriminant axis scores based on the reference regular BoW algorithm, with a quadratic kernel. We have achieved a 65.4% overall accuracy on test datasets, which signifies a 16.3% with our proposed algorithm. Note that, for the quadratic kernel function, it is not possible to retrieve the weights of parameters.

6. Conclusion

We have developed a new type of SURF features based object recognition algorithm, which we refer as Ordered Minimum Distance Bag-of-Words (omidBoW), using an extended set of histograms. This algorithm is evaluated in the context of classifying low resolution aerial object patches, particularly for drones, bird and background. Our proposed approach uses additional sets of histograms compared to regular BoW algorithm, which labels visual words with additional criteria, rather than only choosing the minimum cluster centroid distance. To the best of our knowledge, this is the first such an attempt in the literature, which

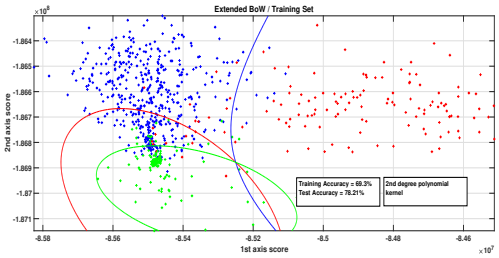

Figure 8. 1st and 2nd discriminant axes scores of training dataset samples and the quadratic boundry lines between classes, for our proposed extended omidBoW approach. (2nd degree polynomial -*quadratic* kernel)

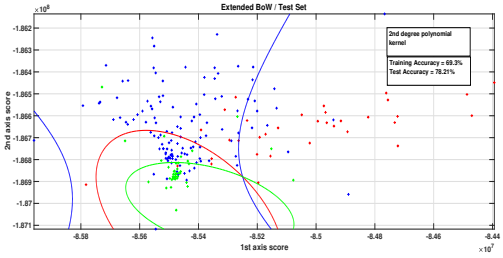

Figure 9. 1st and 2nd discriminant axes scores of test dataset samples and the quadratic boundry lines between classes, for our proposed extended omidBoW approach. (2nd degree polynomial -*quadratic* kernel)

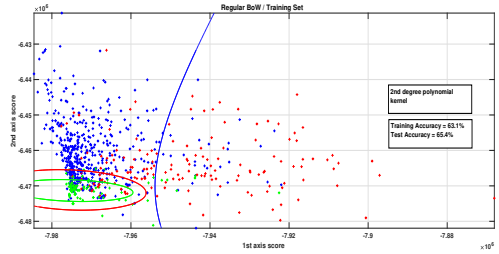

Figure 10. 1st and 2nd discriminant axes scores of training dataset samples and the quadratic boundry lines between classes, for regular BoW approach. (2nd degree polynomial -*quadratic* kernel)

follows the presented approach. Based on our experiments, we have shown that up to 16% percent improvement can be achieved compared to regular BoW approach. In addition to its accuracy, it also offers an extended interpretation of characteristics of visual words which may be more informative. We believe this algorithm can be much more efficient and robust in different context, hence as a future work, we plan to test it in different configurations.

Acknowledgements

This work is supported by French Ministry of Defence.

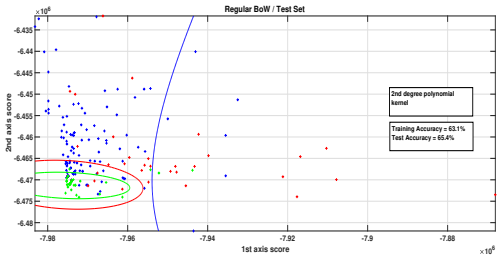

Figure 11. 1st and 2nd discriminant axes scores of test dataset samples and the quadratic boundry lines between classes, for regular BoW approach. (2nd degree polynomial -*quadratic* kernel)

References

[1] J. Atanbori, W. Duan, J. Murray, K. Appiah, and P. Dickinson. Automatic classification of flying bird species using computer vision techniques. *Pattern Recognition Letters*, 81:53–62, 2016.

[2] S. Balakrishnama and A. Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18, 1998.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer vision–ECCV 2006*, pages 404–417, 2006.

[4] I. Colomina and P. Molina. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92:79–97, 2014.

[5] F. Gökçe, G. Üçoluk, E. Şahin, and S. Kalkan. Vision-based detection and distance estimation of micro unmanned aerial vehicles. *Sensors*, 15(9):23805–23846, 2015.

[6] A. J. Izenman. Linear discriminant analysis. In *Modern multivariate statistical techniques*, pages 237–280. Springer, 2013.

[7] L. Kovács and Á. Utasi. Shape-and-motion-fused multiple flying target recognition and tracking. In *SPIE Defense, Security, and Sensing*, pages 769605–769605. International Society for Optics and Photonics, 2010.

[8] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, pages 41–48. IEEE, 1999.

[9] A. Rozantsev, V. Lepetit, and P. Fua. Flying objects detection from a single moving camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4128–4136, 2015.

[10] D. Schmitt and N. McCoy. Object classification and localization using surf descriptors. *CS 229 Final Projects*, pages 1–5, 2011.

[11] P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis. Linear discriminant analysis. In *Robust Data Mining*, pages 27–33. Springer, 2013.