Efficient VLAD in Fine–Grained Image Categorization

ISSN (e) 2520-7393
ISSN (p) 2521-5027
Received on 3rd Nov, 2017
Revised on 20th Dec, 2017
www.estirj.com

Irshad Rahim Memon[1], Khalid Hussain Mohammadani[2*], Sufyan Ali Memon[3], Mumtaz Aziz Kazi[4], Moazzam Ali Bhutto[5]

[1,2,4]Department of Electrical Engineering Isra University, Hyderabad, Pakistan.
[3]Department of Mechanical, Aerospace and Nuclear engineering, Ulsan Institute of science and technology, Ulsan, South Korea.
[5]Department of Information and Communication Engineering, South China university of Technology, China.

Abstract: In this paper, we concentrate the issue of fine-grained images categorization, which is a great deal more helpful in genuine applications than fundamental image classification and we have tried to overcome their problems and to achieve better performance using our proposed technique that is more efficient in terms of memory usage and technical complications as compare to older techniques that having more memory usage and technical complications. For most of datasets Bag of words model and fisher kernel model techniques were used to achieve better performance. In our paper technique used is combination of dense sift and hog descriptors and collectively used with VLAD and finally used by linear SVM classifier that gives better performance and less memory usage and having less technical complications of our taken datasets as compared to previous ones. The two most popular datasets taken are Caltech building dataset and Oxford flowers dataset. For Caltech Building dataset and Oxford flowers dataset we have achieved accuracy up to 87% and 80 % respectively along with best mean accuracy point results through our proposed Efficient VLAD (EVLAD) technique which are better than previous researchers state of art results.

Keywords: Vector local aggregated descriptor, Dense scale-invariant feature transform, Histogram of oriented gradients, mean accuracy point, Support vector machine & Fine-grained image categorization.

1. Introduction

Fine-grained images are the key for some applications like inquiry by images, yet it is extremely testing since it needs to observe the unobtrusive distinction between classes while managing the shortage of training data[1]. As defined in [2], fine-grained picture arrangement alludes to the errand of characterizing articles that have a place with a similar essential level class. Not at all like fundamental level order,fine-grained classification regularly postures difficulties to even profoundly level for recognizing which sub-class a picture having a place with. It is extremely useful for genuine recovery applications. You can envision that you are strolling in a garden, and finding an obscure flower or tree, what will you do? In the event that we have a fine-grained picture recovery framework, then you can counsel the class and inclination about the flower or tree by a moment photograph.Since neighborhood representations are imperviousness to halfway occlusions and jumble and shirking of some preparatory strides in all-encompassing strategies, for example, foundation subtraction and target following, most investigates on protest or image classification depend on nearby components.

These days image-based classification frameworks are accomplishing better and better execution because of expansive datasets and neural systems. In this paper, rather than centering on the errand of characterizing the same number of as various articles, we research the issue of perceiving a huge number of classes inside one classification, for our situation, flowers, and Buiding dataset.Such assignment is called fine-grained classification. There has been progressing in extending the arrangement of the fine-grained domain we have information for, which now incorporates e.g.birds, auto vehicles, aircraft's, buildings, flowers, leaves. In this paper, we concentrate on Flowers and buildings fine-grained arrangement task. For people, we can utilize diverse highlights of flowers and buildings to recognize distinctive species; for the case, we can utilize color, shape, size, and notice data to enable us to settle on a superior choice. Be that as it may, for PCs, the main data they can get is from the info image, which expects us to well plan visual highlights to portray the flowers and buildings clearly.In our paper, we have focused on VLAD in spite of older techniques.

The previous researchers have used sift and hog together along with Bow and we know that Bow consumes more memory usage and its final vector have more size as compared to VLAD[3].We have used the technique of combining dense sift and hog initially and then we got VLAD vector that will be used by SVM classifier.The other parts of this paper are fashioned in the following manner. Section 2 describes literature survey in which different research scholars work will be discussed related to used datasets and about importance and framework of VLAD. Section 3 describes Experiments and discussions in which datasets and results will be discussed separately and section 4 will be the conclusion of this paper.

*Correspondence: khalid.mohammadani@gmail.com

2. Related Work

As our focus is on two popular datasets Caltech building and Oxford flowers dataset so we have extracted other research scholar's work especially related to our datasets. We have also extracted framework of the important parameter of our proposed technique VLAD that will be discussed with help of researcher based definitions and mathematical equations. Authors in [4] have taken flowers dataset that has full 1360 picture dataset comprising of 80 images for each of 17 classifications they have utilized 40 images, 20 train images and 20 test images for every class. They have chipped away at the execution of shading and shape includes independently and acquired 73.7% and 71.8%, of separately with both having 800 clusters. They have accomplished best execution by joining shape and shading with αs =1 and αc =1. In the wake of joining they have obtained 81.3% accuracy. They have additionally contrasted the classifier with a gauge calculation utilizing RGB shading histograms processed for each 10x10 pixel district in an image furthermore, got 55.7% standard execution. BOW utilized alongside standard calculation.

In [5] authors have done the calculation for consequently dividing flowers in shading photos is utilized where 13 flower classes with an aggregate of 753 images are taken. The execution is measured by figuring a cover score, P, between the ground truth division and the division, got by the calculation. It can be seen that taking in an image specific frontal area conveyance significantly enhances the division. In beginning division, they have gotten. The cover of 95% for 62% of the pictures and this is expanded to 77% with 75% of the pictures achieving this point after a solitary emphasis Authors in [6] have examined to what degree mixes of components can enhance classification execution on an extensive dataset of comparative classes. To this end, they have acquainted with 103 class flower dataset. They have registered four distinct components for the flowers, each portraying diverse perspectives, in particular, the nearby shape/surface, the state of the limit, the general spatial appropriation of petals, and the shading. They joined the components utilizing a numerous bit structure with an SVM classifier. Comes about demonstrates that taking in the ideal part blend of different components incomprehensibly enhances the execution, from 55.1% for the best single element to 72.8% for the mix of all elements.

In [7] authors have examined the issue of altogether contrasting two expansive methodologies for measuring picture comparability global versus local features. They have surveyed their execution as the image gathering scales up to more than 11,000 images with more than 6,300 families. They display their outcomes on three datasets with various measurements, including two new difficult datasets. Additionally, they display another calculation to consequently decide the quantity of families in the accumulation with promising outcomes. The global approach speaks to every image by one element descriptor figured from the entire image. The local approach speaks to every picture by an arrangement of local highlighted descriptors registered at some intriguing focuses in the image. CD/DVD amusement covers dataset comprising of 11,431 images and 6,361 families, Caltech structures dataset with 250 images and 50 families, and Oxford structures dataset with 272 images and 11 families. They thought about various sizes of codebooks 1K, 5K, 10K, 25K, and 50K visual words. They additionally think about two variations of BoW: Raw, Tf-idf.BoW. BoW performs best on Caltech and oxford sets while staying inside 10% on the recreations subsets Sift, HOG, and Gist are not reasonable for this undertaking and give much more regrettable outcomes. Creators have prescribed Bag-of-words technique is more the appealing than local features includes as it gives equivalent if worse outcomes while requiring altogether less capacity.

It is a decent possibility for further review. Authors in [8] have displayed a vast scale object recovery framework. The client supplies an inquiry question by selecting an area of an inquiry image, and the framework gives back a positioned rundown of images that contain a similar object, recovered from a huge corpus. They show the versatility and execution of their framework on a dataset of more than 1 million images slithered from the photo sharing site, Flickr [9], utilizing Oxford points of interest as inquiries. They have assessed the execution of their spatial re-positioning on the entire recovery framework for all datasets and look at the impacts of the distinctive change sorts demonstrate the mAP in the wake of applying diverse spatial positioning techniques to the main 100, 200, 400 and 800 returned by the initial filtering stage, for the 5K dataset, utilizing a 20K vocabulary and a 1M vocabulary separately. Contrasting the diverse change sorts on the 5K dataset. Utilizing a 20K visual vocabulary they acquired mAP utilizing Bag of words as 0.385 and by utilizing a 1M visual vocabulary they got mAP utilizing Bag of words as 0.618. In [10] creators have proposed the VLAD representation. They have suggested that The VLAD is the connection of the d-dimensional vectors Vi and is in this manner Kd dimensional. Its primary edge work appears in equation 01

$$Vi \;=\sum\nolimits_{xt:NN(xt)=i} xt - \;\; \mu i \qquad \text{Eq.01}$$

Each local descriptor xt relates to its closest visual word NN(xt) in the codebook. For each codeword μi, the differences $xt - \;\; \mu i$ of the vectors xt assigned to μi are collected. Authors in [11] have proposed a novel Fine-grained Dictionary Learning (FDL) strategy for picture classification. To take in a high-quality discriminating dictionary, three sorts of multi-sub-word references, i.e., class-particular dictionaries, universal dictionary and family-particular word references, are at the same time revealed. Here, class-particular word references and all-inclusive dictionary separately display the examples for each class and the examples regardless of any class. Family-particular dictionaries can help to uncover the common examples between different image classes, by filling the hole between the examples in class-particular word references and all-inclusive dictionary. The reliance among image classes is uncovered by the mutual family-particular dictionary and a typical family-particular word reference can be doled out to a few classes to speak to their lingering. At last, the most segregating family particular word reference for each class is distinguished by limiting the scanty reproduction blunder. Broad investigations were conducted on distinctive generally utilized datasets for image classification. They have also used Oxford flowers17 dataset and by comparing with another scientist result they found FDL gives almost 60% accuracy which is better than other scientist accuracy results. In [12] Authors have proposed a recognition and segmentation algorithm for the reason of fine-grained recognition. The algorithm in the first place distinguishes low-level areas that could conceivably belong to the object and then performs a full-object segmentation through propagation. They tried the algorithm on various benchmark datasets for the fine-grained arrangement. It outflanks all the known state of art methods in oxford flower datasets, once in a while by as much as 11%. It enhances the execution of our benchmark calculation by 3-4%, reliably on all datasets. They additionally watched more than a 4% change in the acknowledgment execution of a testing largescale Oxford flower 17 datasets. Authors in [13] presented three heterogeneous parameters for co-occurrence called color-CoHOG, CoHED, and CoHD, individually. Each heterogeneous parameter is assessed on the INRIA individual dataset and the Oxford 17 class flowers datasets. The result comes in favor of CoHOG for INRIA person dataset and CoHED is best for the Oxford flowers datasets. In [14] [15] authors have also worked on Oxford flowers 17 datasets and their classification accuracy is almost 5% better from [5] but none of the authors have found maP. In [16] researchers have proposed another CNN architecture depend long rectangular kernels to help in redressing moving screen bends from single-images. They have also worked on Caltech building dataset and found 75% accuracy.

In [17] authors have used color-sift method on Caltech building dataset and got 3% less result from [7]. In [18] authors have proposed most discriminative and describing descriptor RCFC by using bit rate scalable descriptor programming. In [19] authors have worked on home identification using mobile object reorganization technique they have found accuracy result initially with no filter and then by applying different filters. It is also better that we have should know regarding dense sift and hog, so we have found some valuable definitions from some reliable sources for dense sift it is stated that the function that has capacity in general identical to running SIFT on a dense grid of locations at a fixed scale and orientation. This sort of highlight descriptors is frequently utilized for object categorization. [20]. The HOG elements are generally used for protest recognition. Hoard decays a picture into little-squared cells, registers a histogram of situated slopes in every cell, standardizes the outcome utilizing a piece shrewd example, and give back a descriptor for every cell. Stacking the cells into a squared picture locale can be utilized as a picture window descriptor for question recognition, for instance by a method for an SVM. [21].

3. Proposed Method

Our proposed model flow chart is shown in figure 01 that gives complete concept of our proposed technique each dataset accuracy and mAP is found separately initially we have taken dense sift and hog descriptors of particular dataset combine them in single vector and used that vector to get 64k vocabulary of dataset as well as their train and test Vlad based vectors and that train and test Vlad based vectors are used by linear SVM classifier to create trained model using train Vlad based vector and label train images and on the basis of that model accuracy and mAP were predicted with help of label test images.

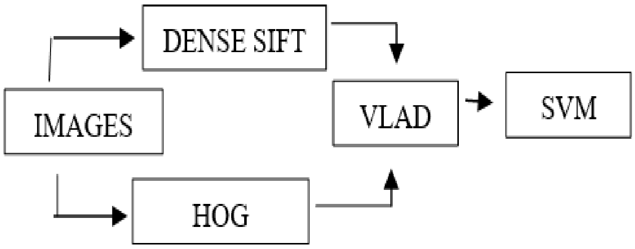

Fig.1. EVLAD Proposed model flowchart

4. Experiments and Discussions

Experiments are performed on two popular datasets Caltech building dataset [22] and Oxford flowers dataset [23].

4.1. Caltech Building dataset

Caltech Building dataset includes 50 different classes of buildings each class has total five images each image shows a different angle of same building that is fruitful in especially training of linear SVM classifier. We will use some of its images for training and some for testing. Each image is focused with different angles and views. After testing efficiency of our used SVM classifier will be recognized along with EVLAD proposed model.

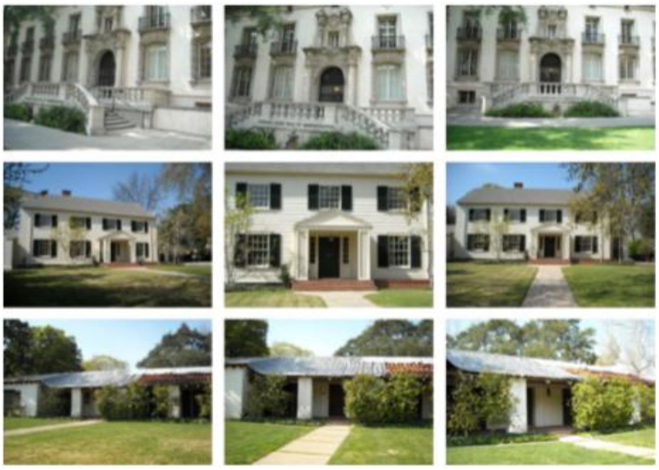

Fig. 2. Test images from caltech dataset. Every line indicates three images for an alternate building taken from various edges and separations [22]

4.2. Oxford Flowers 17 dataset

Oxford flowers dataset consist of 17 classes of flowers each class has 80 images. Each image shows different angles of flowers image that is fruitful in especially training of linear SVM Classifier. We will use some of its images for training and some for testing. Each image is focused with different angles and views. After testing efficiency of our used SVM classifier will be recognized along with EVLAD proposed model.

Fig. 3. Test images from Oxford flowers set. Every line indicates three images for an alternate flower taken from various edges and separations [23]

4.3 Discussions

We have founded performance and mAP of each dataset with help of linear SVM classifier by taking their 30% and 70% training images respectively. For Caltech building dataset, we have obtained 78% accuracy and 0.8163 mAP when 30% training images were selected, and we have obtained 87% accuracy and 0.9007 mAP when 70 training images were selected. Each image is being observed with SIFT and HOG descriptor along with VLAD applied that forms a complete vector of images further their obtained results are shown in table1 and table2 respectively.

Dataset	Accuracy	mAP
Caltech building	**78.1095%**	**0.8163**

Table 1: Accuracy and mAP when 30 % of training images chosen from Caltech building dataset

Dataset	Accuracy	mAP
Caltech building	**87.4172%**	**0.9007**

Table 2: Accuracy and mAP when 70 % of training images chosen from Caltech building dataset

For Oxford flowers dataset, we have obtained almost 72% accuracy and 0.7083 mAP when 30% training images were selected, and we have obtained almost 81% accuracy and 0.8049 mAP when 70% training images were selected. Each image is being observed with SIFT and HOG descriptor along with VLAD applied that forms a complete vector of images further their obtained results are shown in table 3 and table 4 respectively

Dataset	Accuracy	mAP
Oxford flowers	**71.5996%**	**0.7083**

Table 3: Accuracy and mAP when 30 % of training images chosen from oxford flowers dataset

Dataset	Accuracy	mAP
Oxford flowers	**80.9412%**	**0.8049**

Table 4: Accuracy and mAP when 70 % of training images chosen from Oxford flowers dataset

Comparing the other researchers work as mentioned in literature review we have finalized a table that shows best efficient results of other researchers and our results collectively. So, our contribution in getting state of art results is shown in table 05 and table 06.

Method	**Accuracy**	**mAP**
Rengarajan et al [16]	75%	---
Amato et al [17]	82%	---
Lin, Jie, et al [18]	84.2%	---
Aly et al [7]	85%	---
Vu and Thanh Minh [19]	85%	---
Caltech building Dataset (EVLAD)	**87.417%**	**0.9007**

Table 05: Comparison of our proposed EVLAD results with other efficient results of Caltech building Dataset

Method	**Accuracy**	**mAP**
Nilsback and Zisserman [5]	72.8%	---
Ito and Cubota [13]	74.8%	---
Nilsback and Zisserman [14]	76.3%	---
Chai, Bicos method [15]	79.4%	---
Angelova et al [12]	80.66%	---
Oxford Flowers Dataset(EVLAD)	**80.941%**	**0.8049**

Table 06: Comparison of our proposed EVLAD results with other efficient results of Oxford Flower datasets.

5. Conclusion

From the results, it has been concluded that other researchers have got accuracy using a lot of mixing of descriptors and Bag of words and fisher kernel vectors that consumes more vector size and more memory, but our proposed efficient VLAD techniques use the only combination of dense sift and hog descriptors with VLAD vector to give almost approx. the same result even they have focused only on accuracy, but we have focused on accuracy as well as mAP. In Caltech building dataset, we have got efficient results as compared to other researchers but in Oxford flowers dataset we lack some behind. As a conclusion, we have proposed a somewhat better technique that can give us better results even with fewer descriptors and vector size that can have less memory size as compared to mixing a lot of descriptors together with old vectors that can have big memory size and can be complicated in image classification procedures.

References

[1] Feng Zhou, Yuanqing Lin; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1124-1133.

[2] Yao, Bangpeng, Gary Bradski, and Li Fei-Fei. "A codebook-free and annotation-free approach for fine-grained image categorization." In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 3466-3473. IEEE, 2012.

[3] Nister, David, and Henrik Stewenius. "Scalable recognition with a vocabulary tree." In Computer vision and pattern recognition, 2006 IEEE computer society conference on, vol. 2, pp. 2161-2168. IEEE, 2006.

[4] Nilsback, M-E., and Andrew Zisserman. "A visual vocabulary for flower classification." In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2, pp. 1447-1454. IEEE, 2006.

[5] Nilback, Maria-Elena, and Andrew Zisserman. "Delving into the Whorl of Flower Segmentation." In BMVC, pp. 1-10. 2007.

[6] Nilsback, Maria-Elena, and Andrew Zisserman. "Automated flower classification over a large number of classes." In Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on, pp. 722-729. IEEE, 2008.

[7] Aly, Mohamed, Peter Welinder, Mario Munich, and Pietro Perona. "Towards automated large scale discovery of image families." In Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on, pp. 9-16. IEEE, 2009.

[8] Philbin, James, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. "Object retrieval with large vocabularies and fast spatial matching." In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pp. 1-8. IEEE, 2007.

[9] http://www.flickr.com/.

[10] Jégou, Hervé, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. "Aggregating local descriptors into a compact image representation." In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 3304-3311. IEEE, 2010.

[11] Shu, Xiangbo, et al. "Image classification with tailored fine-grained dictionaries." IEEE Transactions on Circuits and Systems for Video Technology 2016.

[12] Angelova, Anelia, and Shenghuo Zhu. "Efficient object detection and segmentation for fine-grained recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[13] Ito, S. and Kubota, S., 2010. Object classification using heterogeneous co-occurrence features. Computer Vision–ECCV 2010, pp.209-222.

[14] Nilsback, Maria-Elena, and Andrew Zisserman. An automatic visual flora: segmentation and classification of flower images. Diss. Oxford University, 2009.

[15] Chai, Yuning, Victor Lempitsky, and Andrew Zisserman. "Bicos: A bi-level co-segmentation method for image classification." In Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2579-2586. IEEE, 2011.

[16] Rengarajan, Vijay, Yogesh Balaji, and A. N. Rajagopalan. "Unrolling the Shutter: CNN to Correct Motion Distortions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[17] Amato, Giuseppe, Fabrizio Falchi, and Paolo Bolettieri. "Recognizing landmarks using automated classification techniques: Evaluation of various visual features." Advances in Multimedia (MMEDIA), 2010 Second International Conferences on. IEEE, 2010

[18] Lin, Jie, et al. "Rate-adaptive compact fisher codes for mobile visual search." IEEE Signal Processing Letters 21.2 pp: 195-198, 2014

[19] Vu, Thanh Minh. "How your phone recognizes your home: An investigation of mobile object recognition." 2016 NCUR 2016

[20] http://www.vlfeat.org/overview/dsift.html.

[21] http://www.vlfeat.org/overview/hog.html.

[22] http:/www.vision.caltech.edu/malaa/datasets/caltech-buildings/.

[23] http://www.robots.ox.ac.uk/~vgg/data/flowers/17/

About Authrs:

Irshad Rahim Memon is Professionaly an Electronic Engineer with more than four year experience in teaching at Undergraduate level.His area of interest includes wireless sensor network, image processing,control Engineering and VLSI design. He obtained his BE in Electronics Engineering from MUET, Jamshoro, Sindh, Pakistan in 2013. He also received M.E Degree in Electronic System Engineering from MUET Jamshoro, Hyderabad, Sindh Pakistan in 2016.Currently, he is serving as a Lecturer in the Department of Electrical Engineering at Isra university Hyderabad, Sindh, Pakistan

Khalid Hussain Mohammadani is a Computer Science Professional with over five-year experience in teaching networking, programming in C, Web development and other CS and Telecommunication courses. His area of interest includes wireless sensor network, mobile ad-hoc network, Localization and Network Coding. He obtained his BS in Telecommunication from University of Sindh, Jamshoro, Sindh, Pakistan in 2008. He also received M. Phil. Degrees in Computer Science from Isra University, Hyderabad, Sindh Pakistan in 2015. At present, he is pursuing Ph.D. in Computer science at Isra University Hyderabad, Sindh, Pakistan.

Dr. Sufyan Ali Memon is professionaly an Electronic Engineer with more than two year experience in teaching at Undergraduate level as Assistant Professor at Department of Electrical Engineering Isra University Hyderabad.His area of interest includes Control Systems Design, Guidance Navigation & Control, Sensor Fusion and Data Fusion. He obtained his PhD in

Electronic System Engineering from Hanyang University, South Korea in 2016.Currently he is doing Post-Doctorate from Department of Mechanical,Aerospace and Nuclear engineering, Ulsan Institute of science and technology, Ulsan , South Korea.

Mumtaz Aziz Kazi is Professionaly an Electrical Engineer with more than seven year experience in teaching at Undergraduate level.His area of interest includes Electrical Power Engineering, and Power Electronics. He obtained his BE in Electrical Engineering from NED, Karachi, Sindh, Pakistan. He also received M.E Degree in Electrical Power Engineering from MUET Jamshoro, Hyderabad, Sindh Pakistan. Currently, he is serving as a Senior Lecturer in the Department of Electrical Engineering at Isra university Hyderabad, Sindh, Pakistan.

Moazzam Ali Bhutto has completed his master in Information and Communication Engineering from South China University of Technolgy. He has earned his two year research experience during his master study. His research area is image processing and speech recognization methods.