International Academy of Engineering and Medical Research, 2017
Volume-2, ISSUE-3
Published Online March 2017 in IAEMR (http://www.iaemr.com)

A Review on Fine Grained Categorization of an Image using Part Proposals

[1]Ms. Dipti S. Borade
PG Student
Department of Computer Engineering
K. K. Wagh Institute of Engineering, Education and Research, Nashik, Maharashtra, India
dipti.borade@gmail.com

[2]Prof. Nitin M. Shahane
Associate Professor
Department of Computer Engineering
K. K. Wagh Institute of Engineering, Education and Research, Nashik, Maharashtra, India
nmshahane@kkwagh.edu.in

***Abstract* - The intent of the image categorization process is to classify the digital image into one of the classes. General image categorization is comparatively easier than fine grained image categorization but it may fail to discriminate objects belonging to same class like birds, cars, plants etc. Fine grained image categorization needs to emphasize on the tiny details that helps to discriminate between similar objects. Many researchers used object /part based methods under strong supervision and weak supervision. The aim is to generate image representation which can be suitable for fine grained categorization. In the new system, object proposals are extracted from input image. From each object proposal, multi-scale part proposals are generated, from which many useful part proposals are selected. A global image representation is generated using selected useful part proposals. The global image representation is then used to train the classifier for image categorization. Application areas are forestry, agriculture, industry and research societies.**

***Keywords* - Fine-grained categorization, feature extraction, part selection**

I. INTRODUCTION

Image categorization refers to classifying images into one of the predefined categories. Fine-grained image categorization is a popular research topic over the past few years. Fine grained image categorization is to categorize objects within an image under some basic level category. For example, to categorize the bird within an image which is of certain class but looks similar to the birds within another class. In fig.1, the Bird has a category Jay, which can be further classified to the subcategories Blue Jay and Green Jay. Thus, fine grained categorization of images is difficult task than general image categorization. In fine grained classification, the emphasis is on the objects which are visually similar to each other. For example, in fig.2 it is easier for a human to differentiate between the dogs and the kangaroo, but the human being may find it difficult to distinguish between two dogs or to decide whether the two dogs are of same category or not. Here is the need of a system which can effectively classify similar images.

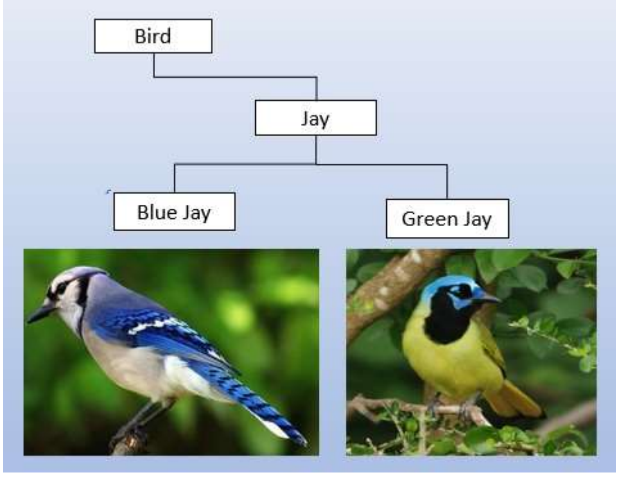

Fig. 1. Fine Grained Categorization Example. Two birds of same category but different subcategory. Blue Jay Vs. Green Jay.

Supervised approach to perform the classification have been used in some or the other kind of annotations in training or testing or in both, training and testing stages. Some of the systems used object bounding box annotations or human in the system loop. Also, object annotations and part annotations have been used in the supervised frameworks. For example, to categorize a bird as an object, its part annotations will be beak, throat, crest, wing, eye, claws nape etc. But it is not always possible to acquire accurate object detectors, part detectors or annotations, also it is expensive and tedious job to acquire the object/ part annotations. Hence it is necessary to classify the fine grained images without using any sort of annotations.

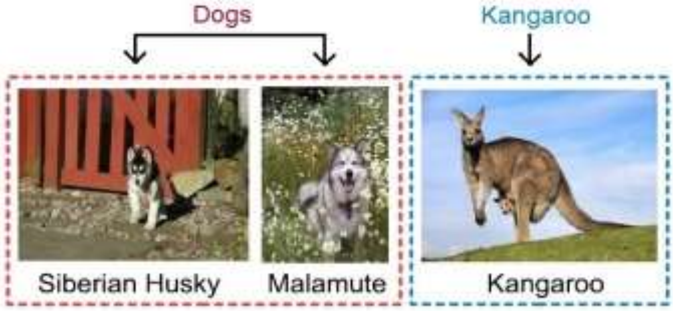

Fig. 2. Fine grained image categorization vs. General image categorization

Traditional general image categorization setups such as scene recognition focus on whole image. Whereas, for fine grained categorization of images, it is necessary to focus on the tiny details within an image such as various parts of an object. The image representations for normal image classification may fail to help categorize images with finer details.

In this weakly supervised system, the aim is to accurately classify the objects using only class labels and without using any other annotations. This work includes three major steps:

- Generate object proposals
- Generate part proposals
- Use multiscale image representation

The first step includes, extracting object proposals i.e. patches from query image. In this work, object proposals are nothing but the regions inside an image having more chances to contain an object of interest. From such object proposals, various parts are drawn at different scales. Among all part proposals, useful parts are selected. These part proposals are then used to get the tiny details to discriminate one entity from another.

The application areas for fine grained image categorization are agriculture, forestry, evaluating the climate changes and forest ecosystem.

II. RELATED WORK

Previous research works includes supervised and semi-supervised approaches for fine grained image classification.

N. Zhang et. al [1] proposed the Deformable Part Descriptors (DPDs), a pose-normalized representation by using Deformable Parts Model (DPM). Many times, the image may contain objects which are not aligned. Two kinds of Deformable Part Descriptors are used. The first DPD, strongly supervised DPD uses the object part annotations for pose normalization. Whereas, weakly supervised DPD uses semantic annotations to train the component correspondence which are then used to get the descriptor with the pose normalized form. More number of supervised part annotations increase the power of DPD but this needs to resolve the issue of self- occlusion.

E. Gavves et. al. [2] proposed the Supervised alignment technique for fine grained categorization. The system proposed by [2] does not contain human interaction. This system do not learn detectors for individual object parts, but instead localize discriminative details by roughly aligning the objects. For alignments, overall shape is considered. These alignments are then used to transfer part annotations from training images to test images (supervised alignment), or to segment the object in a number of regions (unsupervised alignment). Fisher vector encoding-a classification oriented encoding method is used for final image representation.

C. Goring et. all [3] proposed a new part detection method for fine grained recognition. A nonparametric label transfer technique is used which is based on transferring part annotations from related training images to an unseen test image. This allows a feature extraction step that focuses on those parts of images where distinctive features are likely to be located. This method allows for missing part annotations, flexible poses as well as changes in viewpoint.

The success of the technique proposed by [2], [3] depends on the initial alignment. In case of misalignments, the system may fail to perform the correct classification.

N. Zhang et. al [4] proposed the fine grained detection system by using the deep convolutional neural networks. This method learns part models and detectors and learns geometric relationship between parts with the objects. It overcomes the need of using annotations in testing stage. It uses the object bounding box annotations only in training phase. Even with such a system, it is hard to acquire correct annotations needed in the training phase and are hard to obtain for image classification.

C. Wah et. al [5] and T. Xiao et. al [6], proposed the systems with the human in the system loop to help discriminate the objects. In [5], the system makes use of the online game 'Bubbles' that disclose the discriminative features used by human beings. This system proposed BubbleBank algorithm to use the bubbles specified by users to improve system recognition performance. The system proposed in [6] is an interactive system. This does not use the experts to annotate the image parts but use the general users to give attributes. The users are provided with various sets of images and a query image. The users judge the query image and classify it to one of the provided sets of images. This is the unified approach which makes the system simpler. The approaches used in [5], [6] have to rely on human annotations. There may be inaccuracies and subjective differences in human perception system. Sometimes, there may be ambiguities in the response provided by the users. This is the bottleneck of the systems with the human in the system loop.

T. Xiao et. al [6] proposed two level attention model viz. object level attention model and part level attention model. This system pipeline uses deep neural network for categorization. The object level attention model extracts the patches at different views and scales. The part level attention model detects local patterns which are most discriminant. Each part of an object of interest is evaluated by using part detectors and are concatenated to represent the whole image. But it requires to deal with the ambiguities in the part level attention. This system does not use annotations and provides the system with weak supervision. Among previous systems, this is the first system in fine grained categorization setup which does not use manual annotations. It needs two Conv models which adds the computational complexity in the system and degrades performance. Also, it does not fully utilize the resultant outcomes from CNN.

M. Simon and Rodner [7], proposed the unsupervised framework without using any bounding box annotations to learn the model. The pre-learned CNN model is used to generate parts from object of interest. Firstly, outputs from all CNN layers generate a pool of parts. Then, useful parts are chosen for categorization. The system considers two ways to choose useful parts: first is, randomly selecting some parts from all parts; and the second way is to draw a small set of parts by taking into account the relation among them. These parts are then concatenated to describe the final image. But, this may lead to the selection of parts from background. The limitation of this framework is the consideration that a single channel is equivalent to an object part. To improve localization accuracy, a combination of channels can be considered.

D. Yoo et. al [8] proposed the framework for image representation by using the combination of two approaches, the CNN and the fisher vector representation. The Multi-scale Pyramid Pooling (MPP) is proposed in [8]. In this system, to represent an image, a scale pyramid is generated for an input image. All the scaled images are fed into a pre-learned CNN and dense CNN activation vectors are extracted. All such CNN activation vectors are combined by multiscale pyramid pooling. MPP encodes the multiscale resized part feature derived from the local features into separate fisher vector. And then aggregates all such fisher vectors to represent the whole image. The sum pooling method is used to form fisher vector.

M. Jaderberg et. al [9], proposed a new module called Spatial Transformer, which explicitly allows the manipulation of data, spatially within the network. This module can also be included into a neural network to provide various spatial transformation capabilities. The action of the spatial transformer depends on individual data entities. The spatial transformer module is a dynamic mechanism. It spatially and actively transform a feature map of an image by producing transformations for each of the input sample. The transformation like scaling, rotations, cropping, as well as non-rigid deformations are performed for the entire feature map or an image. A limitation of this architecture is that, the number of parallel spatial transformers used in the system limits the number of objects/parts that the network can model.

III. SYSTEM ARCHITECTURE

Fig. 3 shows the System Block Diagram.

The working of the system is elaborated in the following steps:

A. Graph based image segmentation

Input query image is viewed as an undirected graph G = (V,E) where V is the set of pixels and E is the set of edges. Weight of an edge is the absolute difference between pixel intensities. Region pairing Function is used for graph based segmentation. Each region is a set of pixels. Output of this region pairing function is a set of regions.

B. Feature Extraction

In this work, features are extracted as proposals at object level. Object Proposal Generation: An object proposal is the patch from an image which is most likely to contain an object.

In Selective Search[10], Hierarchical Grouping method is used to generate object proposals. For each neighbouring region pair, a similarity measure is calculated. Regions with highest similarity are grouped iteratively.

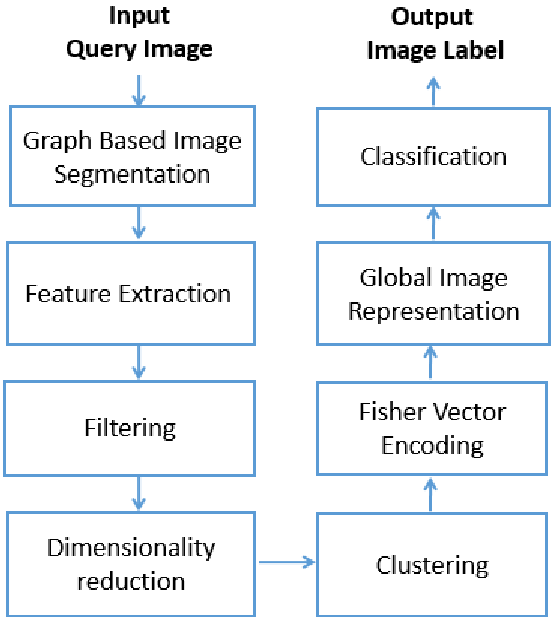

Fig. 3. System Block Diagram.

C. Filtering

Part Proposal Generation: Part proposal is the subregion within an object proposal. In Filtering, part proposals are generated by using the Convolution and Pooling layers on object proposal.

Input to the first convolution layer is the 2D array of pixel values representing an object proposal. Multiscale parts are generated for an object proposal. Hence, the MultiMax Pooling (MMP) technique is used to collaborate the information from all the spatial locations. The system use upto five convolutional layers. The receptive fields of part can highly overlap each other. Fifth convolutional layer conv5 generates multi-scale parts.

D. Optimization

Principal Component Analysis (PCA) technique is used for dimensionality reduction. Feature vectors of parts are reduced by PCA. All the parts are clustered using k-means clustering algorithm.

E. Multiscale Representation of an Image

All the part proposals are encoded by Fisher Vector[11] with Gaussian Mixture Model (GMM). Each Gaussian corresponds to a cluster. Parts generated at different scales must be represent as a whole image to help distinguishing the similar images. Fisher vectors are the state-of-the-art model-dependent but class image representation. Hence, the FV can be successfully used in much applications such as Image classification. It is found out through survey that using FV based approach, the accuracy is increased and less memory is required through approach. In this system, Fisher Vector Encoding is used to encode the part proposals at different scales. An input image comprise a set of object proposals. Each object proposal consists of a set of part proposals at different scales. Fisher vector for each part proposal is concatenated to represent the final image. This representation is used for further classification.

F. Classification

A binary SVM is used to classify the query image. Since this is the multi-class classification problem, one-vs-the-rest strategy is used.

IV. Conclusion

Image Categorization has been a popular research topic from last few years. Fine grained classification needs more attention on discriminative parts between visually similar images. Through survey, we found that both supervised and semi-supervised approaches are used for categorization. Some of the methods reviewed in this paper use annotations to perform classification task but in many cases, these are not easily available. Part based methods can be used to improve the accuracy of image classification. From the experiments carried out in the studied papers, we can say that using part information, system will give better results and help improve categorization accuracy.

REFERENCES

[1] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction, in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Dec. 2013, pp. 729736.

[2] E. Gavves, B. Fernando, C. G. M. Snoek, A. W. M. Smeulders, and T. Tuytelaars, Fine-grained categorization by alignments, in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 17131720.

[3] C. Goring, E. Rodner, A. Freytag, and J. Denzler, Nonparametric part transfer for fine-grained recognition, in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 24892496.

[4] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, Part-based R-CNNs for fine-grained category detection, in Proc. 13th Eur. Conf. Comput. Vis., 2014, pp. 834849.

[5] C. Wah, G. Van Horn, S. Branson, S. Maji, P. Perona, and S. Belongie, Similarity comparisons for interactive fine-grained categorization, in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 859866.

[6] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 842850.

[7] M. Simon and E. Rodner, Neural activation constellations: Unsupervised part model discovery with convolutional networks, in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 11431151.

[8] D. Yoo, S. Park, J.-Y. Lee, and I. S. Kweon. (2014). Fisher kernel for deep neural activations. [Online]. Available: http://arxiv.org/ abs/1412.1628.

[9] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, Spatial transformer networks, in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 20082016.

[10] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, Selective search for object recognition, Int. J. Comput. Vis., vol. 104, no. 2, pp. 154171, Apr. 2013.

[11] J. Snchez, F. Perronnin, T. Mensink, and J. Verbeek, Image classification with the Fisher vector: Theory and practice, Int. J. Comput. Vis., vol. 105, no. 3, pp. 222245, 2013.

[12] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, and M. N. Do, Weakly supervised fine-grained categorization with part-based image representation, IEEE Trans. on Image Processing, vol. 25, no. 4, pp. 17131725, 2016.

Authors' Profiles

Ms. Dipti S. Borade received the B. E. degree in Computer Engineering from Savitribai Phule Pune University, Maharashtra, India. Currently pursuing M. E. degree in Computer Engineering from Savitribai Phule Pune University. Her current research interest includes digital image processing, data mining and information retrieval.

Prof. Nitin M. Shahane Associate Professor in Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik. His current research interests include digital signal processing, digital image processing, pattern recognition, machine learning, data mining and mathematical modelling.