Ensemble Architecture for Improved Image Classification

A. ShubhaRao[1(✉)] and K. Mahantesh[2]

[1] Research Scholar, Department of ECE, SJB Institute of Technology, Bangalore, India
mail2shugar@gmail.com

[2] Associate Professor, Department of ECE, SJB Institute of Technology, Bangalore, India

Abstract. Image classification is one of the major research areas, to meet the needs of reliable and automatic image annotation system. Deep learning techniques have proved to be the solution for most of computer vision problems, with its self-learning ability and non-linear architecture it is able to learn the optimal parameters for the model. Transfer learning based pre-trained models are often used for its flexibility and to achieve good performance with lesser training time. The proposed Ensemble architecture is based on three pre-trained models - Vgg16, Inceptionv3 and Resnet50 Model, the features extracted are merged together using various methods and analyzed on Caltech-101 and Caltech-256 dataset. The conducted empirical study clearly shows that the proposed ensemble model with its merged features outperforms the performance of individual model with greater discriminating ability and with improved accuracy.

Keywords: Caltech-101 · Caltech-256 · InceptionV3 · Resnet50 · Vgg16

1 Introduction

Deep learning had led to greater advancement in the field Artificial Intelligence [1]. Deep learning though sufficient enough, it requires huge amount of data for training and takes too much of time to train the model from the scratch. As mentioned by the authors of VGG model, it took them almost 3 weeks to train the model. Transfer learning comes as a solution, wherein learned parameters (weights) from a pre-trained model are borrowed and used to initialize the model, so that the model gains a kick start when used. Transfer learning aids in achieving better results just within a few minutes [2].

Convolution neural networks (CNN) are basic architecture which captures the major features from an image, by gaining knowledge of unique features which identifies a class [3]. Various kinds of convolution operation like transposed convolution, dilated convolution, depth-wise separable convolution, spatial separable convolution, each with distinguish function capable of extracting a significant feature [4]. Maxpool layers follow up the Convolution layers to reduce the dimension of the data. The data is flattened and passed onto fully connected neural network before making the final prediction. A generalized model of CNN architecture is shown in Fig. 1.

D. S. Guru et al. (Eds.): ICCR 2021, CCIS 1697, pp. 190–199, 2022.
https://doi.org/10.1007/978-3-031-22405-8_15

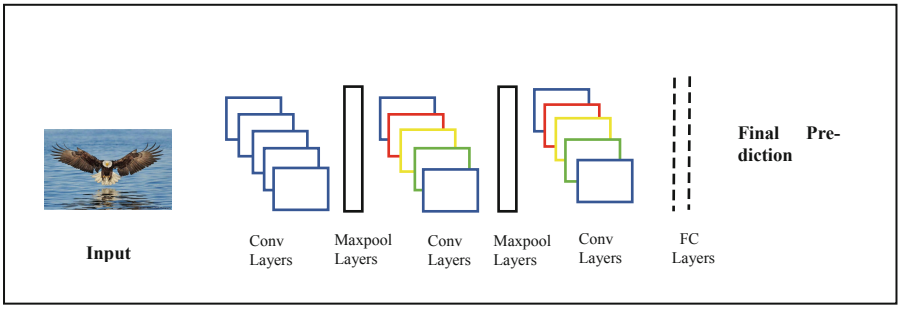

Fig. 1. Generalized CNN Architecture

2 Literature Review

A graph based feature fusion based methodology for image retrieval where outliers are eliminated using Three Degree Binary Graph (TDBG) which is a greedy algorithm, the proposed methodology was tested on publicly available datasets [5]. A method to classify natural scenes which is able to utilize most of the text and character presence in them since text based segmentation methods are more effective. It detector combined with 4 different kind of header is able to classify both at character and text level [6]. A graph based method which concentrates even on local structural regions of the image through hard samples called Graph-based Reasoning Attention Pooling with Curriculum Design (GRAP-CD) is proposed for content based image retrieval [7]. A thorough analysis of various text identification methods like convolutional neural network, maximally stable extreme regions, LSTM for text and character recognition from a scene is presented [8].

A recurrent network based architecture for scene text classification where fixed width, rotation with multi ratio bounding boxes are utilized later proposed sequential regions are further analyzed for textual lines [9]. Mask R-CNN has achieved immense success in the field of object detection yet considerably fails when multiple object instances are present and contains large text case. To overcome, a MLP based decoder which is able to detect and propose compact masks for multiple instances based on shape is proposed. The method shows a significant improvement on five benchmark dataset [10]. Detecting text from a scene has drawn huge attention from various researchers, but success can only be visualized with respect to horizontally and vertically oriented texts. To detect arbitrary and curved texts, a combined architecture based on Proposal Feature Attention Module (PFAM) and One-to-Many Training Scheme (OTMS) is designed which eliminates ambiguity and detects effective feature based on the proposals [11].

To promote comparative diagnostic reading in medical imaging to detect and classify normal and abnormal features separately from images a neural network based architecture is proposed. It classifies the images based on semantic component present and the generated synthesized combined vector [12]. K-nearest neighbor algorithm to get the most To diagnose the lung cancer based on CT images, pre-trained models Vgg16 and Resnet are used to fetch the nearest images for patients. Furthermore the features fetched from Vgg16 are passed onto relatable image [13]. A Fuzzy C means clustering a unsupervised method is used to segment MRI images based on spatial information. The method

is able to locate the clusters even in the presence of noise, without affecting the underlying correlation [14]. With the advancement in cloud computing and cloud storage, the data is encrypted to ensure security. A method which performs effective image matching on encrypted data called Similarity Image Matching (SESIM) is proposed [15].

3 Deep Learning Model

3.1 VGG16 – OxfordNet

Visual Geometry Group from Oxford developed VGG16 a Convolutional Neural Network based model, it won 1st runner in the ILSVRC (ImageNet) Challenge of the year 2014 [16]. It is one of simplest yet effective architecture ever proposed to extract the features from the image. The architecture consists several blocks of 3*3 convolutional layers followed by max-pooling layers, increasing the depth gradually from 64, 128, 256 to 512. At the top of the stack, the data is passed to series of Fully Connected Layers (Dense), before making the final prediction [17]. The architecture of VGG16 is shown in Fig. 2.

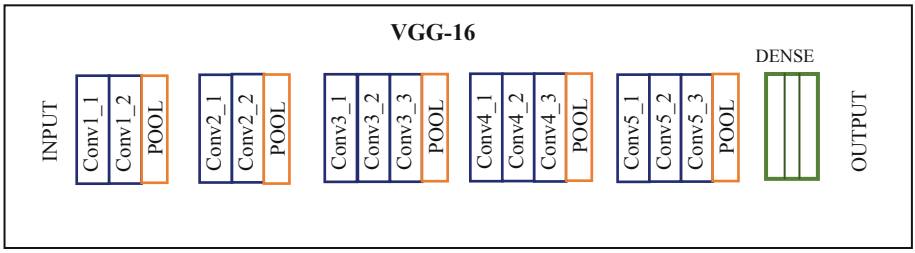

Fig. 2. Architecture of VGG16

3.2 Inceptionv3 – GoogLeNet

The model which won the 2014 ILSVRC challenge was InceptionNet. Inception-v3 also called as the second generation Inception, is an architecture proposed by the authors to improve the efficiency of the classifier along with reduced computational complexity of the model. The major architectural changes proposed 3 different kind of inception blocks - Factorized convolutional block- single 5*5 conv was replaced by two 3*3 conv, replacing 3*3 conv by 1*3 and 3*1 conv, the idea was to make the architecture not only deeper but wide enough to capture the spread out features [18]. The authors also added batch normalization layer into auxiliary classifiers, along with label smoothing. The building blocks of Inception Model are shown in Fig. 3.

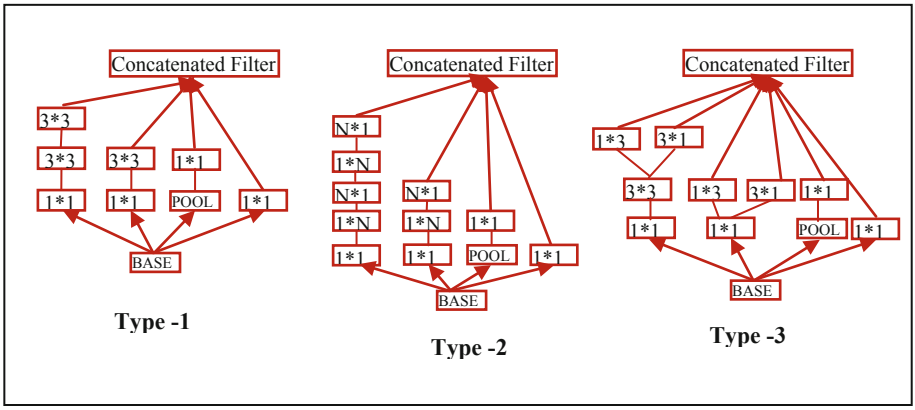

Fig. 3. Architecture Blocks of Inception-v3

3.3 ResNet – 50

ResNet is the architecture which won the 2015 ILSVRC of image classification. Resnet was mainly designed to address the persistent problem of exploding/ vanishing gradient whenever the network is deeper [19]. The issue addressed by adding Residual blocks in the architecture which is a skip connection between the layers. To further reduce the complexity of the model 1*1 – 3*3 – 1*1 conv blocks were added as sandwich layers. The architecture of ResNet-50 is shown in Fig. 4.

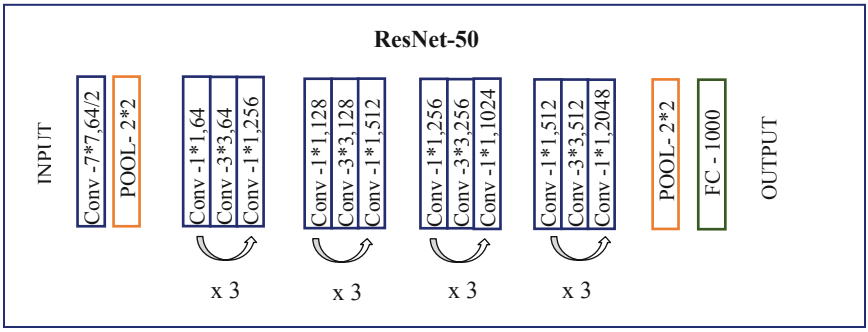

Fig. 4. Architecture of ResNet-50

4 The Proposed Ensemble Model

For the proposed ensemble approach Vgg16, Inceptionv2 and ResNet50 pre-trained models with top is used as feature extractor, the feature vectors are merged together with merging layer, followed by two fully connected layers of 1000 neurons, before making the final prediction. Before going to final ensemble, vgg+inception was used

to analyze the behavior of various feature merging techniques. To merge the feature extracted from different models merging layers like - Add, Concat, Max, Min, Subtract were analyzed based on which adding the features together was chosen as most ideal for the final ensemble model. The outline of the proposed model is shown in Fig. 5.

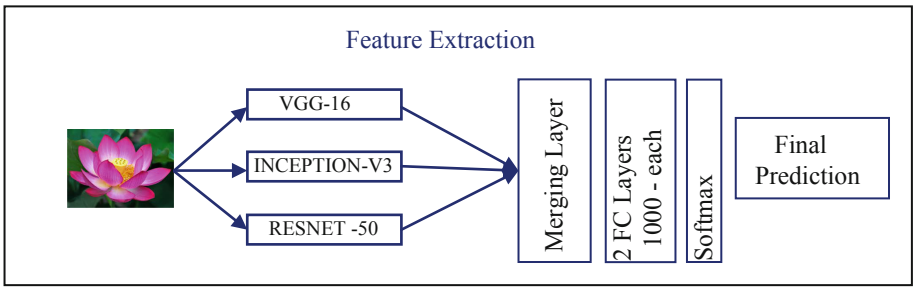

Fig. 5. Architecture of proposed ensemble model

Algorithm:

1. Load the dataset, create train and test dataframes.
2. Create the model with specified optimizer, compile the model. Optimizer= 'Adagrad'
3. Initialize the model with weights pre-trained of 'Imagenet'.
4. Read the images from the respective dataframe, apply data augmentation.
5. Pass the data into their corresponding pre-processing function, as required by the pre-trained model.
6. Collect the feature vectors from different models.
7. Merge the feature vectors to create the final feature vector,

$$V_{Vgg16} + V_{Inceptionv2} + V_{Resnet50} = V_{Merged\ Feature\ Vector} \quad (1)$$

8. Pass the merged features to fully connected layers.
9. Use Softmax to make the final prediction.
10. Fit the model for train data, evaluate on test data.

5 Results

Caltech-101 and Caltech-256 object category dataset is chosen for the study and analysis of the proposed ensemble model. Caltech-256 is one highest object category dataset next to ImageNet with around 30,000 images [20]. The analysis made with different merging techniques on Caltech-101 and Caltech-256 dataset, using 2 pre-trained models it can be clearly seen from the results obtained that ADD – adding the feature vectors together will increase weightage of the particular feature detected resulting in better performance as shown in Table 1. The performance of the models is measured and compared in terms of accuracy (%).

Table 1. Analysis of various feature merging techniques on Caltech-101 and Caltech-256 dataset on Vgg16+Inceptionv3 Model

Methods	**Caltech-101 30-Train**	**Caltech-256 30-Train**
Vgg16+Inceptionv3 ----- Concat	65	53
Vgg16+Inceptionv3 ----- Add	70	55
Vgg16+Inceptionv3 ----- Subtract	63	47
Vgg16+Inceptionv3 ----- Max	60	45
Vgg16+Inceptionv3 ----- Multiply	25	12
Vgg16+Inceptionv3 ----- Average	62	51
Vgg16+Inceptionv3 ----- Min	20	10

The selected merging technique is applied in the proposed ensemble model and evaluated for 15-Train, 30-Train, as opposed to deep learning papers where the authors choose to split the train data with 60% of weightage. Comparative analysis of Caltech-101, Caltech-256 dataset results with previous work is tabulated in Table 2 and Table 3 respectively. It is evident from the result obtained that proposed ensemble model which is a combination of 3 different pre-trained models and the merged features, outweighs the performance of all the previous work.

Table 2. Comparative analysis of Proposed Ensemble Model on Caltech-101 result with previous work

Methods	**Caltech-101 15-Train**	**Caltech-101 30-Train**
Shape matching [21]	45	–
Pyramid match kernels [22]	49.5	58.2
Discriminative nearest neighbour [23]	59	66

(*continued*)

Table 2. (*continued*)

Methods	**Caltech-101 15-Train**	**Caltech-101 30-Train**
Local naïve bayes nearest neighbour [24]	47.8	55.2
Sparse localized features [25]	33	41
Relevance based classification [26]	–	43.8
Gaussian mixture models [27]	–	72.3
VGG-16 [28]	66	78.42
Inceptionv3 [29]	64	67
Proposed Ensemble Model	**73.11**	**79.23**

Table 3. Comparative analysis of Proposed Ensemble Model on Caltech-256 result with previous work

Methods	**Caltech-256 15-Train**	**Caltech-256 30-Train**
Learning dictionary [30]	30.35	36.22
Sparse spatial coding [31]	30.59	37.08
Discriminative coding for object classification [32]	28	30
Local naïve bayes classifier [24]	33.5	40.1
Caltech Institute classification [33]	28.3	34.1
Combined image descriptors [34]	–	33.6
VGG-16 [28]	51	57.57
Inceptionv3 [29]	58	59
Proposed Ensemble Model	**60**	**62**

6 Conclusion

The Proposed Ensemble Model outperforms the state-of the-art techniques like VGG16, Inception and ResNet50. Study also justifies the choice of these models for merging as each of the models has its own unique nature and methodology to extract feature from an image. The model performs well even with smaller train data, in comparison to other research work where 60% of train data is used to training the model. The idea of merging the features has powered the model with discriminating ability to further increase the accuracy achieved as against the individual models. The research also justifies the fact that with the aid of transfer learning, innovative, efficient, simple models can be designed. However, the proposed model fluctuates with validation data, henceforth concentrating on stabilizing the model still poses itself as a challenge.

References

1. Bayhan, E., Ozkan, Z., Namdar, M., Basgumus, A.: Deep learning based object detection and recognition of unmanned aerial vehicles. In: 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1–5 (2021). https://doi.org/10.1109/HORA52670.2021.9461279
2. Saeed, M., Nagdi, M., Rosman, B., Ali, H.H.S.M.: Deep reinforcement learning for robotic hand manipulation. In 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), pp. 1–5, (2021). https://doi.org/10.1109/ICCCEEE49695.2021.9429619
3. Chiba, S., Sasaoka, H.: Basic study for transfer learning for autonomous driving in car race of model car. In 2021 6th International Conference on Business and Industrial Research (ICBIR), pp. 138–141 (2021). https://doi.org/10.1109/ICBIR52339.2021.9465856
4. Kumar, D., Kukreja, V.: N-CNN based transfer learning method for classification of powdery mildew wheat disease. In: 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 707–710 (2021). https://doi.org/10.1109/ESCI50559.2021.9396972
5. Lao, G., Liu, S., Tan, C., Wang, Y., Li, G., Xu, L., Feng, L., Wang, F.: Three degree binary graph and shortest edge clustering for re-ranking in multi-feature image retrieval. J. Vis. Commun. Image Represent. **80**. https://doi.org/10.1016/j.jvcir.2021.103282 (2021)
6. Wu, D., Hu, X., Xie, Z., Li, H., Ali, U., Lu, H.: Text detection by jointly learning character and word regions. In Lladós, J., Lopresti, D., Uchida, S. (eds.) Document Analysis and Recognition – ICDAR 2021. Lecture Notes in Computer Science, vol. 12821. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86549-8_20
7. Zhu, X., Wang, H., Liu, P., Yang, Z., Qian, J.: Graph-based reasoning attention pooling with curriculum design for content-based image retrieval. Image Vis. Comput. **115**, 104289, ISSN 0262-8856 (2021). https://doi.org/10.1016/j.imavis.2021.104289
8. Gupta, M., et al.: Analysis of text identification techniques using scene text and optical character recognition. IJCVIP **11**(4), 39–62 (2021). https://doi.org/10.4018/IJCVIP.2021100104
9. Zou, B., Yang, W., Liu, S., Jiang, L.: Multi-oriented scene text detection by fixed-width multi-ratio rotation anchors, Comput. & Electr. Eng. **95**, 107428, ISSN 0045-7906 (2021). https://doi.org/10.1016/j.compeleceng.2021.107428
10. Qin, X., Zhou, Y., Guo, Y., Wu, D., Tian, Z., Jiang, N., Wang, H., Wang, W.: Mask is all you need: rethinking mask R-CNN for dense and arbitrary-shaped scene text detection. CoRR abs/2109.03426 (2021)

11. Guo, Y., Zhou, Y., Qin, X., Wang, W.: Which and where to focus: a simple yet accurate framework for arbitrary-shaped nearby text detection in scene images. In: Artificial Neural Networks and Machine Learning – ICANN 2021. Lecture Notes in Computer Science, vol. 12895. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86383-8_22
12. Kobayashi, K., Hataya, R., Kurose, Y., Miyake, M., Takahashi, M., Nakagawa, A., Harada, T., Hamamoto, R.: Decomposing normal and abnormal features of medical images for content-based image retrieval of glioma imaging. Med. Image Anal. **74**,102227, ISSN 1361-8415 (2021). https://doi.org/10.1016/j.media.2021.102227
13. Rajasenbagam, T, Jeyanthi, S.: Semantic content-based image retrieval system using deep learning model for lung cancer CT images. J. Med. Imaging Health Inform. **11**(10), 2675–2682(8) (2021). https://doi.org/10.1166/jmihi.2021.3859
14. Kamarujjaman, Maitra, M., Chakraborty, S.: A novel spatial FCM-based method for brain MRI image segmentation in the presence of noise and inhomogeneity. In: Maji, A.K., Saha, G., Das S., Basu S., Tavares J.M.R.S. (eds.) Proceedings of the International Conference on Computing and Communication Systems. Lecture Notes in Networks and Systems, vol. 170. Springer, Singapore (2021). https://doi.org/10.1007/978-981-33-4084-8_37
15. Janani, T., Brindha, M.: Secure Similar Image Matching (SESIM): an improved privacy preserving image retrieval protocol over encrypted cloud database. IEEE Trans. Multimed. (2021). https://doi.org/10.1109/TMM.2021.3107681
16. Gu, J., Yu, P., Lu, X., Ding, W.: Leaf species recognition based on VGG16 networks and transfer learning. In: 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2021, pp. 2189–2193 (2021). https://doi.org/10.1109/IAEAC50856.2021.9390789
17. Aung, H., Bobkov, A.V., Tun, N.L.: Face detection in real time live video using yolo algorithm based on Vgg16 convolutional neural network. In: 2021 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), pp. 697–702 (2021). https://doi.org/10.1109/ICIEAM51226.2021.9446291
18. Singprayoon, S., Supratid, S.: Effects of number and position of auxiliary networks used in inception convolutional neural network on object recognition. In: 2021 9th International Electrical Engineering Congress (iEECON), pp. 452–455 (2021). https://doi.org/10.1109/iEECON51072.2021.9440065
19. Wang, Y., Zhao, Z., He, J., Zhu, Y., Wei, X.: A method of vehicle flow training and detection based on ResNet50 with CenterNet method. In: International Conference on Communications, Information System and Computer Engineering (CISCE), pp. 335–339 (2021). https://doi.org/10.1109/CISCE52179.2021.9446012
20. Mahantesh, K., Shubha Rao, A.: Content based image retrieval - Inspired by computer vision & deep learning techniques. In 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), pp. 371–377 (2019). https://doi.org/10.1109/ICEECCOT46775.2019.9114610
21. Berg, T.L., Berg, A.C., Malik, J.: Shape matching and object recognition using low distortion correspondence. In: IEEE CVPR, 1, 26–33 (2005)
22. Grauman, K., Darell, T.: Pyramid match kernels: discriminative classification with sets of image features. Technical report MIT-CSAIL-TR-2006-020 (2006)
23. Maire, M., Malik, J., Zhang, H., Berg A.C.: SVM-KNN: discriminative nearest neighbor classification for visual category recognition. In: IEEE-CVPR, 2:2126–2136 (2006)
24. McCann, S., Lowe, D.G. Local naive bayes nearest neighbor for image classification. In: IEEE-CVPR, pp. 3650–3656 (2012)
25. Mutch, J., Lowe, D.G.: Muticlass object recognition with sparse, localized features. IEEE CVPR, 1:11 18 (2006)

26. German Gonzalez EnginTuretkenFethallahBenmansour Roberto Rigamonti, Vincent Lepetit. On the relevance of sparsity for image classification. Comput. Vis. Image Underst. **125**, 115127 (2014)
27. Mahantesh, K., Aradhya, V.N.M., Niranjan, S.K.: An impact of complex hybrid color space in image segmentation. In: Recent Advances in Intelligent Informatics. Advances in Intelligent Systems and Computing, Springer, vol. 235, pp. 73–83 (2014). https://doi.org/10.1007/978-3-319-01778-5
28. AS Rao K Mahantesh 2021 Learning semantic features for classifying very large image datasets using convolution neural network SN Computer Science 2 3 1 9 https://doi.org/10.1007/s42979-021-00589-6
29. Rao, A.S., Mahantesh, K.: Image Classification based on Inception-v3 and a mixture of Handcrafted Features, Lecture Notes in Electrical Engineering (LNEE), Springer book series, [Accepted manuscript - Article in Press], Series/7818, ISSN: 1876–1100 (2021)
30. Zhang, Y.-J., Liu, B.-D., Wang, Y.-X.: Learning dictionary on manifolds for image classification. Pattern Recognit. **46**, 1879–1890 (2012)
31. Vieira, A.W., Campos, M.F., Oliveira, G.L., Nascimento, E.R.. Sparse spatial coding: a novel approach for efficient and accurate object recognition. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 2592–2598 (2012)
32. K Balakrishna 2020 WSN-based information dissemination for optimizing irrigation through prescriptive farming Int. J. Agric. Environ. Inf. Syst. (IJAEIS) 11 4 41 54 https://doi.org/10.4018/IJAEIS.2020100103
33. Holub, A., Griffin, G., Perona, P.: Caltech 256 object category dataset. Technical Report, California Institute of Technology (2007)
34. Banerji, S., Sinha, A., Liu, C.: New image descriptors based on color, texture, shape, and wavelets for object and scene image classification. Neurocomputing **117**, 173–185 (2013)