Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549686

HYBRID MODEL FOR IMAGE CLASSIFICATION: BRIDGING THE GAP WITH GLOBAL FEATURE DESCRIPTOR

Shubha Rao A[a,1] and Dr. Mahantesh K [b]

[a] Research Scholar, Department of ECE, SJB Institute of Technology, Bangalore, India
[b] Associate Professor, Department of ECE, SJB Institute of Technology, Bangalore, India

Abstract. With the progressive advancement in Artificial Intelligence, Deep learning techniques have stood out with its remarkable performance in various computer vision applications like image classification, object detection, segmentation. A feature fusion methodology which bridges the gap between the machine and deep learning technique, captivating the benefits of both the field is proposed. Machine level features – color, texture, shape which describes an image are fused with the higher level semantic features extracted by state-of-the-art technique. Convolutional Neural Network (CNN) based Vgg16 architecture is used to as feature extractor which are later fused with machine level features. The performance of the algorithm is measured on Caltech-101and Caltech-256 object category datasets using various machine learning classifiers. The proposed methodology with its powerful fused descriptor outperforms the state of the art complicated models.

Keywords. *Feature Fusion, Vgg16, Machine Learning, Caltech101, Caltech256.*

INTRODUCTION

All the advancement witnessed in the field of Artificial intelligence and robotics is majorly intervened with the progress in deep learning techniques. Deep learning is a technique inspired by the human nervous system, is a subset of machine learning [1]. Difference in process of feature extraction adopted by techniques is what distinguishes them majorly. Machine Learning involves major human interference in determining the filters for the feature extraction vs. Deep learning is self-sufficient enough to figure out the best filters for any required application [2]. Machine learning algorithms mainly, learn by seeing examples – supervised learning, learn by practice – unsupervised learning, learn with little help – semi-supervised learning [3,4]. Deep learning algorithms learn the best filters for a particular task – Convolutional Neural Network (CNN), learn to benefit from the gained knowledge in the path – Recurrent Neural Network (RNN) [5]. In the today's digital era, automatic annotation of image (classification) is about to become a major perquisite for all applications. The proposed methodology of future fusion is an attempt to gain the benefits from both the techniques, for the task of image classification. The visualization in major difference between the two approaches of AI can be seen in Figure 1.

[1] Shubha Rao A, Research Scholar, Department of ECE, SJB Institute of Technology, Bangalore, India; E-mail: mail2shugar@gmail.com.

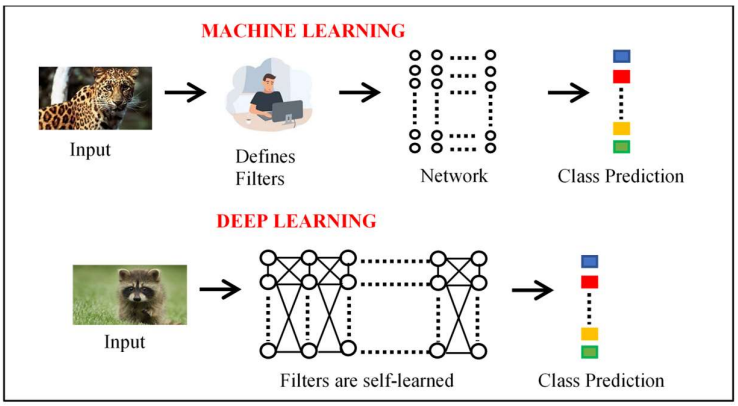

Figure 1. Machine Learning vs. Deep Learning

LITERATURE REVIEW

A transfer learning based deep learning model inspired by YOLO to classify large Kannada Scene Individual Character (KSIC) has achieved 8% greater test accuracy in comparison with other models [6]. Heterogeneous Future fusion technique is proposed to make effective use of text present in images by creating an embedded vector, the framework is composed of Multi-model Complementary Fusion (MCF), Cross Model Guided pooling (CGP), and Relative Caption-aware Consistency (RCC) to combine the local features of both text and image [7]. Another method to improve the efficiency of image classification with attention mechanisms of neural network by utilizing both global feature channels along with local attention which mimics the human visual attention mechanism helps in faster convergence of the model. [8]. Improved cross-modal image retrieval which eliminates the ambiguity that exits between the query image and the bias of target database by getting the clue in the form of user feedback and database re-ranking accordingly is proposed [9]. Fine tuning based transfer learning approach when applied on smaller dataset suffers from bottleneck issue at time of training. The issue is solved by adapting a regularization technique where outer layers weights are constrained by using starting point as reference (SPAR) and by selecting a subset of features using attention mechanism [10].

A SIRe method which preserves the local structure of the image for classification by using skip and residual connections and with interlaced auto encoders merged with base CNN model is proposed [11]. An interactive content based image retrieval system where visual saliency maps of the retrieved images are made visible to the end users and the relevant user feedback merged together. The proposed method is tested on MS-COCO dataset displays an improvement in accuracy [12]. To effectively merge the modification text along with the query image while comparing with the target image is proposed employing a Joint Prediction Module (JPM) which can be adopted with any existing architecture to benefit the implicit knowledge of the query image [13]. To bridge the inevitable gap that exits in representation of data between the sketches and photos a Domain Aware Squeeze and Excitation (DASE) network is proposed. The network reduces the maximum distance between intra class variability by adding a multiplicative Euclidean distance which is represented by loss function Multiplicative Euclidean Margin Softmax (MEMS) ultimately creating a diverse feature space for effective image retrieval [14].

A quantum cuckoo search optimization (QCSO) a better optimization technique to take the CNNs to the next level where it is capable to recognize the localized parts of the images for their respective class is proposed [15]. A dynamic attention mechanism with multiple heads is used to recognize the various channels, which is capable of extracting the multiple local features of image. The extracted features are merged and the weighted sum of which is used for image retrieval [16]. A StyleGAN based approach to disentangle the orthogonal feature vectors and spread them into sparse latent feature space. The features are later selected and assigned preferences depending on the requirement of attributes present in the query image [17]. A feature fusion based method to detect long text present in the images by using enlarged receptive fields within the atrous convolution module is proposed. The features extracted at different scales are fused again to avoid the inconsistency and invariance across the feature pyramid [18].

HYBRID MODEL: THE PROPOSED METHODOLOGY

In order to utilize the benefit of both the world (ML and DL) for the task of image classification a hybrid model is been proposed. Since the efficiency of image classification majorly narrows down to the quality of features which are fed to classifiers, a global feature descriptor based on feature fusion technique is projected. At first a set of hand crafted low-level basic features which define the core of image - color, texture and shape are extracted individually. To match with human competency in recognizing images a set of semantic features are extracted using Convolutional Neural Network. The handpicked and semantic features are fused together to create global feature descriptor, the efficiency of which is tested on various machine learning classifiers. A detailed description of method of feature extraction, feature fusion and classification techniques is outlined in further subsections.

3.1 TYPES OF FEATURES EXTRACTED

Color: One of the major features which contribute the most in any of image processing task is color. Color Histogram is a statistical way to analyze the distribution of the color of an image [19]. The image is converted into HSV color space before computing the histogram which utilizes the advantage of having better human perception. Histograms can be considered as bar graph of number of colors present to number of pixels in an image. The histograms are normalized and flatted to fetch the color feature vector (f_{color}).

Texture: The feature which gives specific spatial information and its relationship among the pixels of an image is texture. Texture gives a fair idea of variation in the smoothness, sharpness of an image which is used extensively for segmentation. Haralick features which basically uses Gray Level Co-occurrence Matrix (GLCM) to average out the features along the given direction for any given region of interest [20]. Given query image is converted into gray scale, extracted haralick features are stored as texture feature vector ($f_{texture}$).

Shape: The feature which is most frequently used for object recognition and its classification is shape. Shape helps in identifying the objects metrics like shape, length and position. One of the simplest ways to identify edges is based on Histogram of Oriented Gradients (HOG) [21]. The image is resized which is a prerequisite for HOG function; the extracted features are flattened before saving as shape feature vector (f_{shape}).

$$||\nabla g|| = \sqrt{\left(\frac{\partial g}{\partial x}\right)^2 + \left(\frac{\partial g}{\partial y}\right)^2} \quad (1)$$

Semantic Features: There is always a difference with manually extracting the low-level features from image to high level semantic features which are perceived by human. CNNs are the best in learning such ideal filters along with automatically fetching the semantic features. VGG16 one of the simplest, most ideal CNN with its stacked layers of convolution layers and gradual increase in the filter size proves to be one of the best CNN architecture for image classification [22]. VGG16 model which is pre-trained using Imagenet dataset is used just before the fully connected layers (top) is used to extract the semantic feature vector (f_{vgg16}). The architecture of vgg16 is shown in Figure 2.

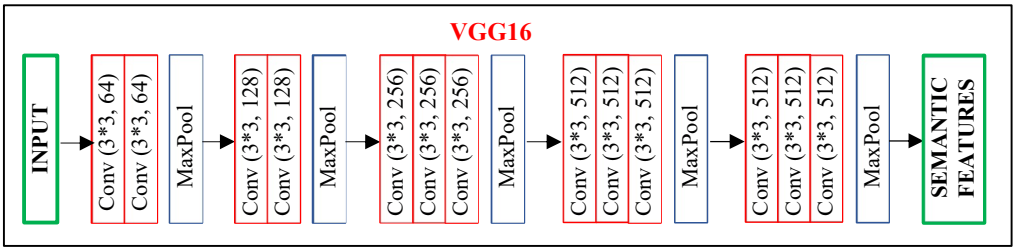

Figure 2. VGG16 architecture used for semantic feature extraction

3.2 GLOBAL FEATURE DESCRIPTOR: BRIDGING THE GAP

Once the various features from the image like the color, texture, shape and semantic features are extracted it is concatenated to create global feature descriptor (f_{global}). A data frame of global feature descriptor is created along with their respective labels (classes) of image, which can be later divided for train and test purpose. To examine the proposed hybrid model, the global features are passed onto various machine learning classifiers and trained. Once the classifiers are trained, the performance will be measured for test data. The detailed architecture of the proposed Hybrid Model is shown in Figure 3.

$$f_{global} = f_{vgg16} \parallel f_{color} \parallel f_{texture} \parallel f_{shape} \quad (2)$$

The various machine learning classifiers which were used for the pragmatic study are as follows:

Logistic Regression: A statistical algorithm is mainly used for classification purpose inspired by the probability to understand the relationship between various variables using sigmoid cost function.

K – Nearest Neighbor: The classifier uses all the given data employing a lazy non-parametric method to classify new data depending on its nearest neighbor.

Random Forest: An ensemble based learning approach by merging (majority) the output of different decision trees to increase the overall performance (classification).

Decision Tree: A tree structure is used recursively to learn to make the right decision which eventually increases the homogeneity of the subset of data.

Support Vector Machine: The non - parametric algorithm classifies the data by finding the best fitting hyper plane with maximized margin to the nearest neighbor point.

Stochastic Gradient Descent: Starting from a random data point the algorithm tries to find a global minima of optimization function depending on the direction of gradient.

$$\theta_j = \theta_j - \alpha(\hat{y} - y) \quad (3)$$

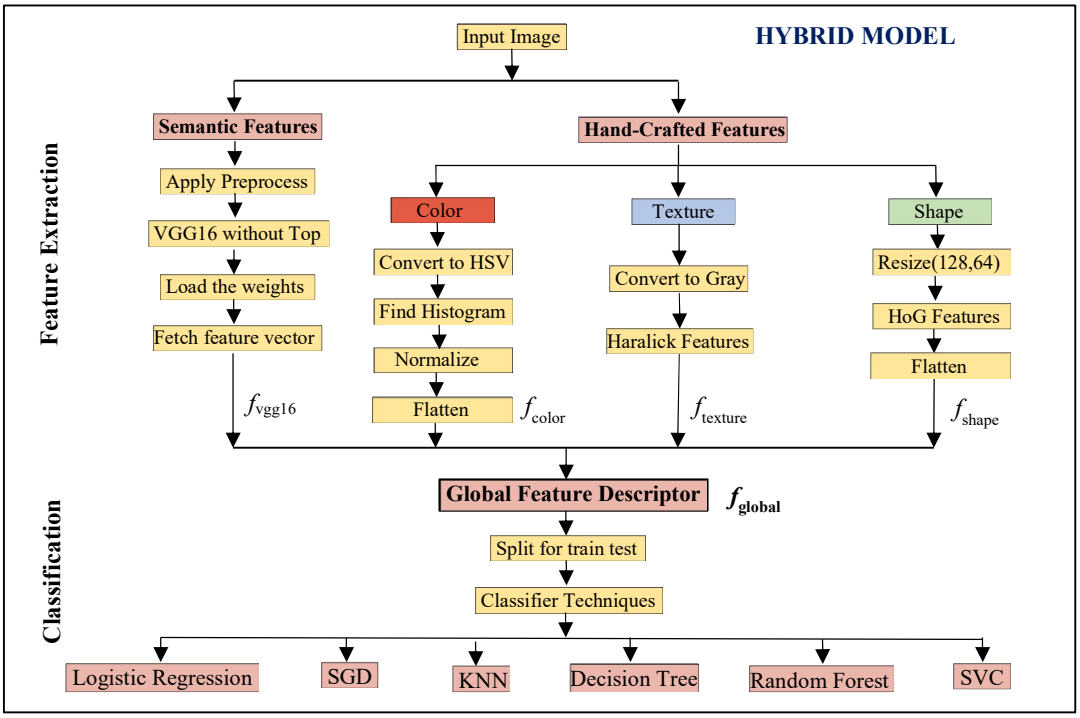

Figure 3. Detailed architecture of Hybrid Model

3.3 RESULT & DISCUSSION

The performance of the Hybrid model is tested of one largest image classification dataset Caltech-101 with 9,146 images and Caltech-256 with 30,607 images [23]. Caltech dataset has higher object category in comparison to other most commonly used MNIST Digital dataset (10 Category), CIFAR (10 and 100 category), COCO (80 object category). While splitting the dataset to measure the efficiency of the model, 30 images from each category is used for training and remaining images are used for testing. The performance of the model is analyzed for various classifier techniques the result of which is tabulated in Table.1. From the table it can be clearly inferred that Logistic Regression outperforms SGD, KNN, Decision Tree, Random Forest and SVC.

The performance of the hybrid model with its global feature descriptor on Caltech-101, Caltech-256 datasets is compared with the previous state-of-the art techniques, the fair analysis of which is shown Table .2 and Table .3 respectively. It can be clearly concluded from the results the proposed methodology outperforms the earlier work which were based on sparse localized features, pyramid match kernels, learned dictionaries and successfully beats even the deep learning based pre-trained models. It makes it worth mentioning the designed architecture proves its efficiency even with smaller sample space (15 Train) on Caltech-101 dataset. The performance for Hybrid model on Caltech-256 surpasses all the earlier work with exception to VIRNet architecture where it bit lesser accuracy. With increased object category the performances is getting dropped due the overlapping of features caused by cross references of information across various level. The performance of the model is measured using accuracy as metric in terms of percentage (%).

Table 1. Comparative analysis of Proposed Hybrid Model performance using various Classifiers

Classifiers	Caltech-101 30-Train	Caltech-256 30-Train
Logistic Regression	**91.37**	**72**
KNN	45.16	18
Random Forrest	82.73	50
Decision Tree	51.3	20.52
SVC	73	63
SGD	87	69

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

Methods Accuracy (%)	Caltech-101 15-Train	Caltech-101 30-Train
Shape matching [24]	45	-
Pyramid match kernels [25]	49.5	58.2
Discriminative nearest neighbour [26]	59	66
Local naïve bayes nearest neighbour [27]	47.8	55.2
Sparse localized features [28]	33	41
Relevance based classification [29]	-	43.8
Gaussian mixture models [30]	-	72.3
VGG-16 [31]	66	78.42
Inceptionv3[32]	64	67
Ensemble Model [33]	73.11	79.23
VIRNet [34]	87.74	91
Proposed Method: Hybrid Model	**88.52**	**91.37**

Table 3. Comparative analysis of Proposed Hybrid model on Caltech-256 result with previous work

Methods Accuracy (%)	Caltech-256 15-Train	Caltech-256 30-Train
Learning dictionary [35]	30.35	36.22
Sparse spatial coding [36]	30.59	37.08

Discriminative coding for object classification [37]	28	30
Local naïve bayes classifier [38]	33.5	40.1
Caltech Institute classification [23]	28.3	34.1
Combined image descriptors [39]	-	33.6
VGG-16 [31]	51	57.57
Inceptionv3[32]	58	59
Ensemble Model [33]	60	62
VIRNet [34]	70.28	74.25
Proposed Method: Hybrid Model	**68**	**72**

CONCLUSION

The proposed Hybrid model with the specially designed global feature descriptor is able to exceed even the performance of current artificial intelligence based neural network architecture. With the successful effort to create one global feature descriptor based on feature fusion which takes the advantage of both the low-level (machine) features and high-level (semantic) feature for image classification. The feature extraction methods are carefully handpicked such that it remains simple, easy to compute and yet efficient enough to deliver expected results. Tackling the problem of feature overlapping across multiple levels with increased object category is a challenge.

The hybrid model clearly demonstrates with innovative ideas it is possible to build a bridge between ML and DL which are rich with features for any computer vision application. As a part of future work the architecture needs to be tested on various other dataset belonging to diverse domain.

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao and S. C. H. Hoi, "Deep Learning for Person Re-identification: A Survey and Outlook," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2021.3054775.

[2] C. -Y. Huang-Fu, C. -H. Liao and J. -Y. Wu, "Comparing the performance of machine learning and deep learning algorithms classifying messages in Facebook learning group," 2021 International Conference on Advanced Learning Technologies (ICALT), 2021, pp. 347-349, doi: 10.1109/ICALT52272.2021.00111.

[3] J. Datta, R. Dasgupta, S. Dasgupta and K. R. Reddy, "Real-Time Threat Detection in UEBA using Unsupervised Learning Algorithms," 2021 5th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), 2021, pp. 1-6, doi: 10.1109/IEMENTech53263.2021.9614848.

[4] K. Patel and H. B. Patel, "A Comparative Analysis of Supervised Machine Learning Algorithm for Agriculture Crop Prediction," 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2021, pp. 1-5, doi: 10.1109/ICECCT52121.2021.9616731.

[5] S. Sundari, E. C. Djamal and A. Wulandari, "Classification of Emotion Based on Electroencephalogram Using Convolutional Neural Networks and Recurrent Neural Networks," 2021 International Conference on Instrumentation, Control, and Automation (ICA), 2021, pp. 213-218, doi: 10.1109/ICA52848.2021.9625702.

[6] S. S. N. Sunilkumar S Manvi, "Transfer Learning based Kannada Text Detection and Recognition in Natural Scene Images", *DE*, pp. 9374-9388, Nov. 2021.

[7] Gangjian Zhang, Shikui Wei, Huaxin Pang, and Yao Zhao. 2021. Heterogeneous Feature Fusion and Cross-modal Alignment for Composed Image Retrieval. Proceedings of the 29th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, pp. 5353–5362. doi:https://doi.org/10.1145/3474085.3475659.

[8] Abraham Montoya Obeso, Jenny Benois-Pineau, Mireya Saraí García Vázquez, Alejandro Álvaro Ramírez Acosta, Visual vs internal attention mechanisms in deep neural networks for image classification and object detection, Pattern Recognition, Volume 123, 2022, 108411, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2021.108411.

[9] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, and Miki Haseyama. 2021. Database-adaptive Re-ranking for Enhancing Cross-modal Image Retrieval. In Proceedings of the 29th ACM International Conference on Multimedia (MM '21). Association for Computing Machinery, New York, NY, USA, pp.3816–3825. doi:https://doi.org/10.1145/3474085.3475681.

[10] Xingjian Li, Haoyi Xiong, Zeyu Chen, Jun Huan, Ji Liu, Cheng-Zhong Xu, and Dejing Dou. 2021. Knowledge Distillation with Attention for Deep Transfer Learning of Convolutional Networks. ACM Trans. Knowl. Discov. Data 16, 3, Article 42 (June 2022). doi:https://doi.org/10.1145/3473912.

[11] Danilo Avola, Luigi Cinque, Alessio Fagioli, Gian Luca Foresti**, SIRe-Networks: Skip Connections over Interlaced Multi-Task Learning and Residual Connections for Structure Preserving Object Classification.** *arXiv - CS - Artificial Intelligence* (IF), doi*: arxiv-2110.02776*. 2021.

[12] Vasu, Bhavan & Hu, Brian & Dong, Bo & Collins, Roddy & Hoogs, Anthony. Explainable, Interactive Content-Based Image Retrieval. Applied AI Letters. 10.1002/ail2.41. (2021).

[13] Yuchen Yang, Min Wang, Wengang Zhou, and Houqiang Li. 2021. Cross-modal Joint Prediction and Alignment for Composed Query Image Retrieval. Proceedings of the 29th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, pp. 3303–3311. doi:https://doi.org/10.1145/3474085.3475483.

[14] Peng Lu, Gao Huang, Hangyu Lin, Wenming Yang, Guodong Guo, and Yanwei Fu. 2021. Domain-Aware SE Network for Sketch-based Image Retrieval with Multiplicative Euclidean Margin Softmax. In Proceedings of the 29th ACM International Conference on Multimedia (MM '21). Association for Computing Machinery, New York, USA, pp. 3418–3426. doi:https://doi.org/10.1145/3474085.3475499.

[15] Sarva Naveen Kumar, Dr. Ch. Sumanth Kumar. (2021). Convolution Neural Networks for Content based Image Retrieval Using Quantum Cuckoo Search Optimization Technique. *Design Engineering*, 5688 - 5701. Retrieved from http://www.thedesignengineering.com/index.php/DE/article/view/5529.

[16] Hui Wu, Min Wang, Wengang Zhou, Houqiang. "Learning Deep Local Features With Multiple Dynamic Attentions for Large-Scale Image Retrieval". In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11416-11425.

[17] A Zaeemzadeh, S Ghadar, B Faieta, et al. Face Image Retrieval with Atrribute Manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[18] Liangjun Wang, Weijie Gu, and Yuhang Ji "Scene text detection with improved receptive field and adaptive feature fusion", Proc. SPIE 11911, 2nd International Conference on Computer Vision, Image, and Deep Learning, 119110V (5 October 2021); https://doi.org/10.1117/12.2604527.

[19] Y. Yankun, D. Xiaoping, C. Wenbo and W. Qiqige, "A Color Histogram Based Large Motion Trend Fusion Algorithm for Vehicle Tracking," in IEEE Access, vol. 9, pp. 83394-83401, 2021, doi: 10.1109/ACCESS.2021.3087904.

[20] Y. Sari, A. R. Baskara and R. Wahyuni, "Classification of Chili Leaf Disease Using the Gray Level Co-occurrence Matrix (GLCM) and the Support Vector Machine (SVM) Methods," 2021 Sixth International Conference on Informatics and Computing (ICIC), 2021, pp. 1-4, doi: 10.1109/ICIC54025.2021.9632920.

[21] M. Kitayama and H. Kiya, "Generation of Gradient-Preserving Images allowing HOG Feature Extraction," 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), 2021, pp. 1-2, doi: 10.1109/ICCE-TW52618.2021.9603248.

[22] J. Duan and X. Liu, "Online Monitoring of Green Pellet Size Distribution in Haze-Degraded Images Based on VGG16-LU-Net and Haze Judgment," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-16, 2021, Art no. 5006316, doi: 10.1109/TIM.2021.3052018.

[23] A. Holub G. Griffin and P. Perona. Caltech 256 object category dataset. Technical Report, California Institute of Technology, 2007.

[24] T. L. Berg A. C. Berg and J. Malik. Shape matching and object recognition using low distortion correspondence. In IEEE CVPR, 1:26–33, 2005.

[25] K. Grauman and Darell. Pyramid match kernels: Discrminative classification with sets of image
features. Technical report MIT-CSAIL-TR-2006-020, 2006.

[26] Maire. M-Malik Hao Zhang, Berg A.C. Svm-knn: Discriminative nearest neighbor classification for
visual category recognition. In IEEE-CVPR, 2:2126–2136, 2006.

[27] Sancho Mc Cann and David G. Lowe. Local naive bayes nearest neighbor for image classification. In IEEE-CVPR, pages 3650–3656, 2012.

[28] Jim Mutch and David G Lowe. Muticlass object recognition with sparse, localized features. IEEE CVPR, 1:11 18, 2006.

[29] German Gonzalez Engin Turetken Fethallah Benmansour Roberto Rigamonti, Vincent Lepetit. On the relevance of sparsity for image classification. Computer Vision and Image Understanding, 125:115127, 2014.

[30] 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. doi.org/10.1007/978-3-319-01778-5 8.

[31] Rao, A.Shubha, Mahantesh, K. Learning Semantic Features for Classifying Very Large Image Datasets Using Convolution Neural Network. *SN COMPUT. SCI.* **2,** 187 (2021). https://doi.org/10.1007/s42979-021-00589-6.

[32] Image Classification based on Inception-v3 and a mixture of Handcrafted Features, Lecture Notes in Electrical Engineering (LNEE), Springer book series, 2021 [Accepted manuscript - Article in Press], Series/7818, ISSN: 1876-1100.

[33] Rao, A.Shubha, Mahantesh, K. Ensemble Model for improved Image Classification. In: The International Conference on Cognition and Recognition (ICCR), Springer book series LNNS [Accepted manuscript - Article in Press], 2021.

[34] Rao, A.Shubha, Mahantesh, K, Vidyashree Nagaraju. VIRNet for Image Retrieval: One for All Top based on Feature Fusion Technique, HCII-2022, Springer [Accepted manuscript], CCIS series, 2022.

[35] Yu-Jin Zhang Bao-Di Liu, Yu-Xiong Wang. Learning dictionary on manifolds for image classification. Pattern Recognition, 46:1879–1890, 2012.

[36] Antonio W. Vieira Mario F. M. Campos Gabriel L. Oliveira, Erickson R. Nascimento. Sparse spatial coding: A novel approach for efficient and accurate object recognition. IEEE International Conference on Robotics and Automation (ICRA), pages 2592–2598, 2012.

[37] Zhang Y Zheng Y Liu B, Wang Y. Discriminant sparse coding for image classification. Proceedings of the 37th International Conference on Acoustics, Speech and Signal Processing, pages 2193–2196, 2012.

[38] Sancho Mc Cann and David G. Lowe. Local naive bayes nearest neighbor for image classification. In IEEE-CVPR, pages 3650–3656, 2012.

[39] Chengjun Liu Sugata Banerji, Atreyee Sinha. New image descriptors based on color, texture, shape, and wavelets for object and scene image classification. Neurocomputing, 117:173–185, 2013.