Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018)
IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1-5386-0965-1

Transfer learning for image classification

Manali Shaha[1] Meenakshi Pawar[2]
Dept. of Electronics and Telecommunication
SVERI, Pandharpur
Maharashtra, India
[1]mmpawar@coe.sveri.ac.in
[2]manalishah1191@gmail.com

***Abstract*—Convolutional neural network (CNN) gained great attention for robust feature extraction and information mining. CNN had been used for variety of applications such as object recognition, image super-resolution, semantic segmentation etc. due to its robust feature extraction and learning mechanism. By keeping constant the baseline learning topology, various CNN architectures were proposed to improve the respective system performance. Among these, AlexNet, VGG16 and VGG19 are the famous CNN architecture introduced for object recognition task. In this paper, we make use of transfer learning to fine-tune the pre-trained network (VGG19) parameters for image classification task. Further, performance of the VGG19 architecture is compared with AlexNet and VGG16. Along with the CNN architectures, we have compared the hybrid learning approach which is comprised of robust feature extraction from CNN architecture followed by support vector machine (SVM) classifier. We have used two state-of-the-art databases namely: GHIM10K and CalTech256 to study the effect of CNN architecture for robust feature extraction. Performance evaluation has been carried out using average recall, precision and F-score. Performance analysis shows that fine-tuned VGG19 architecture outperforms the other CNN and hybrid learning approach for image classification task.**

***Index Terms*—Image classification, AlexNet, VGG16, VGG19, CalTech256, GHIM10K**

I. Introduction

Research in image classification witnessed the evolution in computer vision algorithm from first order moments to hand crafted features to end-to-end machine learning approaches to improve the classification accuracy. This evolution was initialized by extracting textural information using first order moments and grey level dependency features. Haralick et al. [1] proposed set of simple texture features. They proposed grey level dependency statistics for texture classification. Further, structural approach for texture information has been proposed by Haralick et al. [2] to incorporate the structural information in texture classification. Manjunath et al. [3] proposed first order moments to extract the texture information with an application to content based image retrieval. However, first order moments are not invariant to scale as well as rotation. To overcome the scale and rotation invariance, Han et al. [4] proposed rotation and scale invariant Gabor filters with application to texture classification. Further, supervised learning proposed by Talbar et al. [5] for texture classification.

However, first order moments are not robust to classify the similar contrast textures. Also, it fails in complex textures. Ojhala et al. [6] proposed local binary patterns (LBP) to extract the local neighbourhood information. LBP has proved its effectiveness in almost all computer vision algorithm because of its simple and computationally efficient implementation. Though, it is unable to incorporate the directional information, variants of LBP were proposed according to the need of application. To extract illumination invariant local features, Tan et al. [7] proposed local ternary patterns (LTP). Further, Murala et al. proposed bank of local operators [8]–[16] with an application to CBIR. Among which, local tetra patterns [14] extract twelve directional information and obtain more robust information. Spherical symmetric 3D LTP proposed by Murala et al. [11] extracts the spatio-temporal information. Not only CBIR but also variety of applications [17], [18] make use of local feature extraction, because of its ease of use. Further, some modified approach proposed by Mayuri et al. [18] for CBIR using combination of local feature descriptor and artificial neural networks.

Even though these operators extract the local information they fail in complex scene or clutter background. Evolutionary, interest point detection methods overcome the drawbacks of local operator and other hand-crafted features. Initially, Harris et al. [19] proposed corner detection algorithm based on pure mathematical theory. However, corner detection method fails with variation of scale. Further, Lowe et al. [20] proposed scale invariant feature transform (SIFT) to detect the scale invariant interest points (SIIP). SIIP followed by histogram of oriented gradients extracts the robust SIFT features. Speeded up robust features (SURF) [21] introduced by Bay et al. to reduce the computational complexity of SIFT. Further, various researchers integrate the interest points detected by SIFT/SURF with another feature descriptor and introduced different robust feature descriptor [22]–[24] for image classification and other computer vision tasks.

Feature descriptor discussed above are witnessed to the growth of the digital image libraries and the evolution of large scale databases. Hand crafted features fail in the large-scale database because of the high intra as well as inter class variation in image categories. However, effective/robust feature descriptor could improve the performance of image classification. In recent years, convolutional neural network had great success in almost all areas of machine learning and computer vision filed. Due to the robustness of CNN feature extraction, researchers make use of it in variety of applications. Initially, Alex et al. [25] proposed an evolutionary

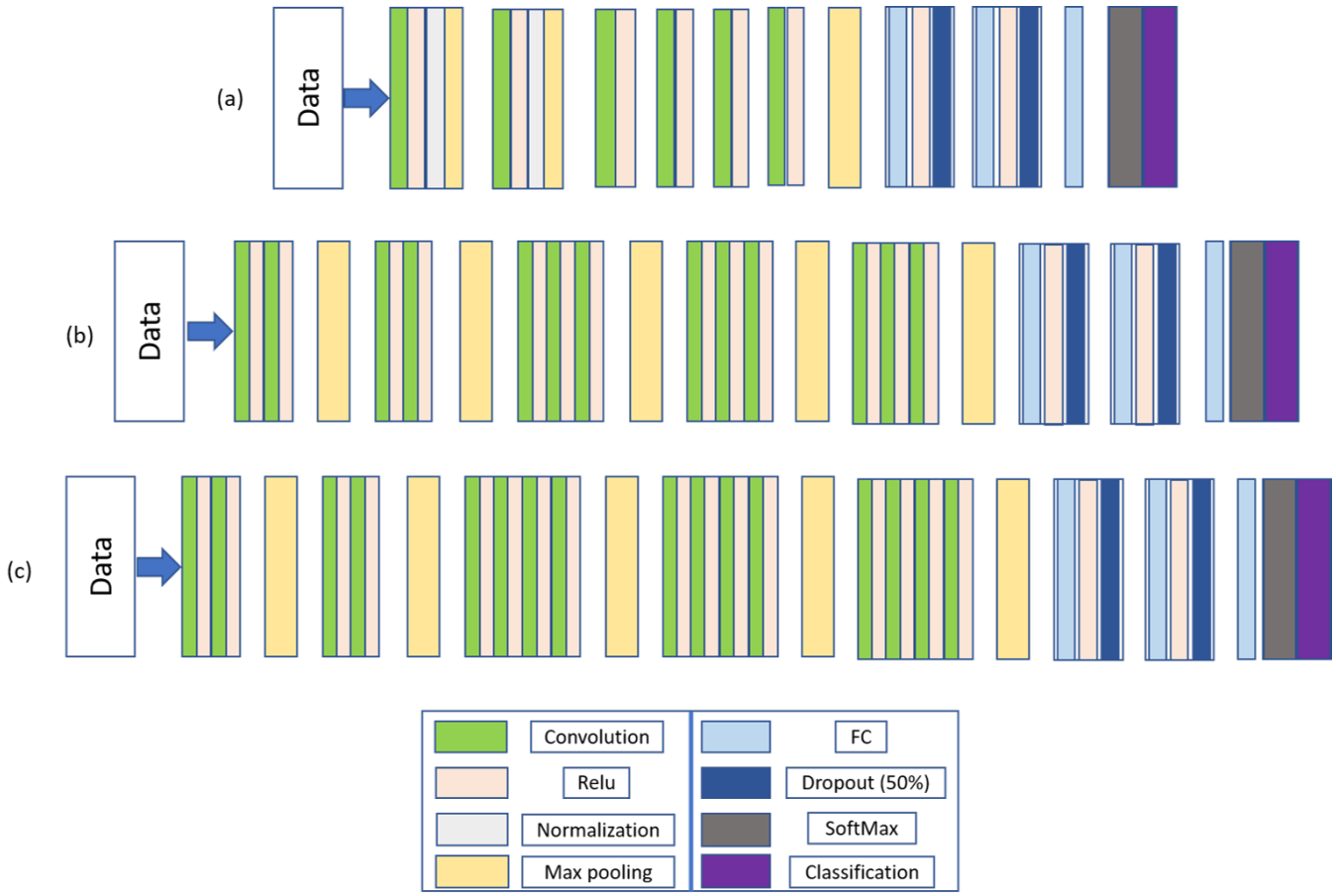

Fig. 1. Existing CNN architectures. (a) AlexNet [25] (b) VGG16 [27] (c) VGG19 [27]

CNN architecture named as AlexNet for object recognition task. The major hurdle in training of CNN is availability of large scale database. However, they used ILSVRC [26] database along with augmentation to train their network. To improve the recognition accuracy further towards the human vision system researchers proposed deeper CNN architectures. Simonyan el al. [27] proposed VGG16 architecture for object recognition task. Improved VGG16 architecture known as VGG19 overcomes the drawbacks of AlexNet and increases the system accuracy.

In this paper, we have fine-tuned the VGG19 architecture over two state-of-the-art databases CalTech256 [28], GHIM10K [29] and analysed the effect of deeper network by comparing the result with AlexNet and VGG16 architectures for image classification task. Also, we have analysed the effect of SVM classifier in conjunction with the extracted features from CNN architectures. Performance evaluation carried out using recall, precision and F-score.

II. Existing CNN architecture

In this section, we have discussed three existing CNN architectures AlexNet [25], VGG16 [27] and VGG19. Initially, AlexNet proposed by Alex et al. [25] to solve the object recognition problem. It was the first try to learn the network parameters for recognition task over very large scale database. AlexNet consist twenty six layers, out of which last two layers are softmax and output layers. Network architecture is divided into three parts. Fig. 1 (a) shows the network architecture. First part of the network consists of two units, each unit comprises { convolution, relu, normalization and pooling layer}. Second part of the network consist of four units, each of them comprises { convolution and pooling layer}. Last part of the network corresponds to the non-linear activation unit which corresponds to the fully connected (FC), relu and drop-out layer. Drop-out layer avoids the over-fitting of the data during training. This repetitive structure of the network adapts the data characteristics and extracts the robust features. Initial filters learnt the low level features.

Accuracy of the CNN architecture highly depends upon the three factors namely: Large scale database, high end computational unit and the network depth. Out of these, requirement of the training database is solved due to availability of the ILSVRC [26] publicly available database. GPU unit can solve the second difficulty. However, last parameter has an uncertainty, because there is no such measure which could set a limit for network depth. Going deeper in the network extracts more complex and robust features. Simonyan el al. [27] proposed VGG16 architecture for object detection task. Fig. 1 (b) shows

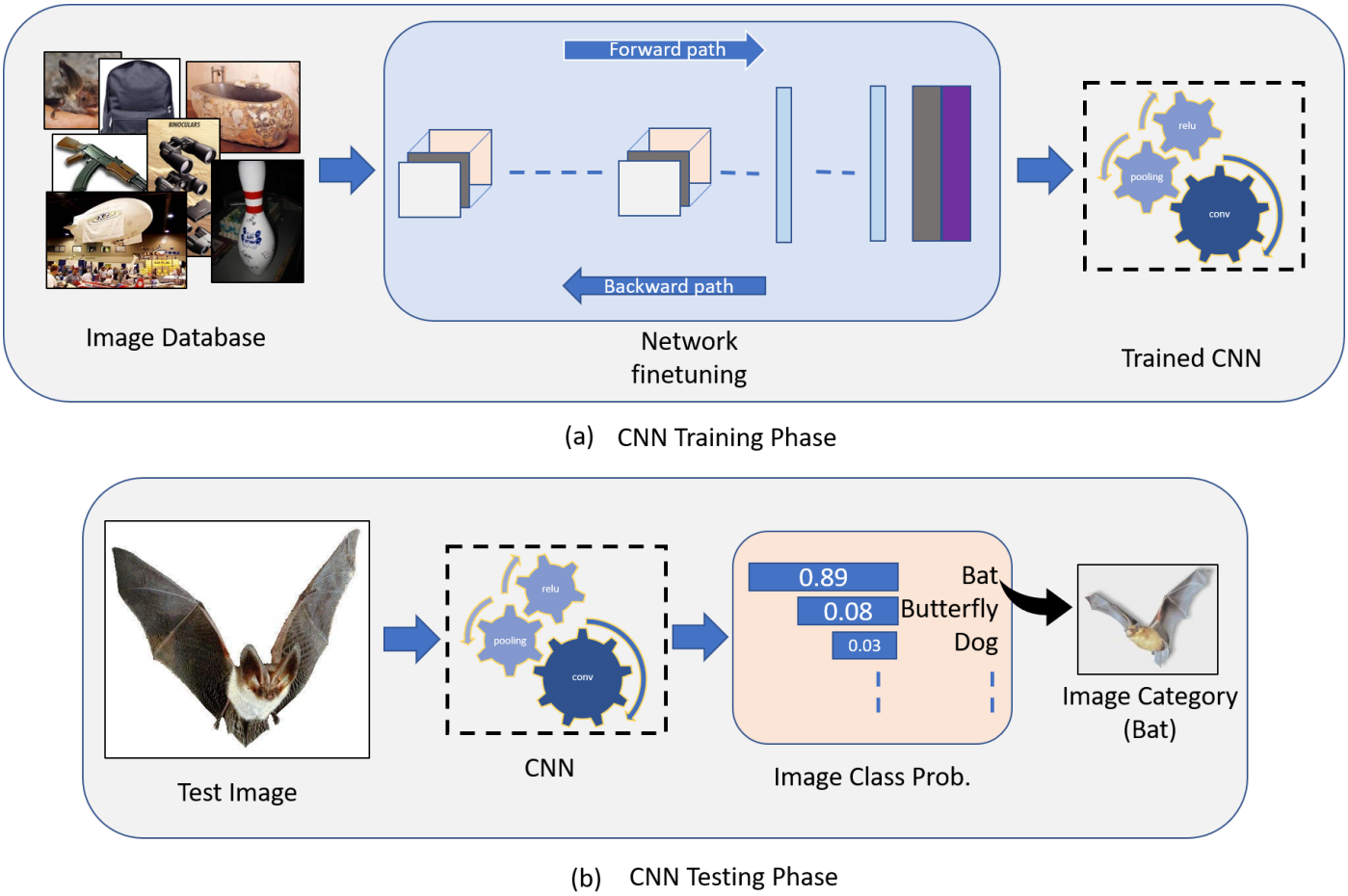

Fig. 2. Proposed system block diagram. (a) CNN transfer learning (fine tuning) (b) CNN testing phase

the network architecture. Unlike AlexNet, VGG16 consists of replicative structure of { convolution, relu and pooling layer}. They increased the number of such network unit to design deeper network. However, [27] considered smaller size receptive window for each convolutional filter as compared to AlexNet. Non-linear activation unit is same as that of AlexNet. Further, more deeper network VGG19 is proposed for the same task (Object detection). VGG19 comprises of some extra convolutional relu units in the middle of the network as compared to VGG16. However, this minute change in the architecture turns into the accuracy enhancement for object recognition task.

III. Proposed Approach

ILSVRC database consist of 22000 categories of objects. It covers almost all the existing object whichever known to a common human being. However, existing networks are trained over 1000 different categories out of 22000. It is quite impossible to learn the network parameters of such huge networks over small scale datasets. However, these parameters can be fine tuned over the small datasets as per the application demands.

In this work, we have fine-tuned the network parameters of the VGG19 over two databases namely: CatlTech256 and GHIM10K. Proposed network is divided into two parts, (1) CNN training phase and (2) CNN testing phase. Fig. 2 shows the proposed system flow. Fig. 2(a) illustrates the CNN training phase in which network parameters of the VGG19 are fine-tuned and trained VGG19 is obtained. However, Fig. 2(b) shows the CNN testing phase which comprises the test image followed by trained VGG19 to estimate the image class probability.

IV. Experimental Results

We have divided our analysis into three experiments. Out of which first two were performed on two state of the art databases namely: GHIM10K and CalTech256.

A. Experiment #1

We have carried this experiment on publicly available GHIM10K database. It consists of 20 classes, each class is having 500 images. Fig. 4 shows the sample images from GHIM10K dataset. In this experiment, we have analysed the performance of VGG19 architecture for image classification task. Fig. 3 illustrates the class wise average { recall, precision

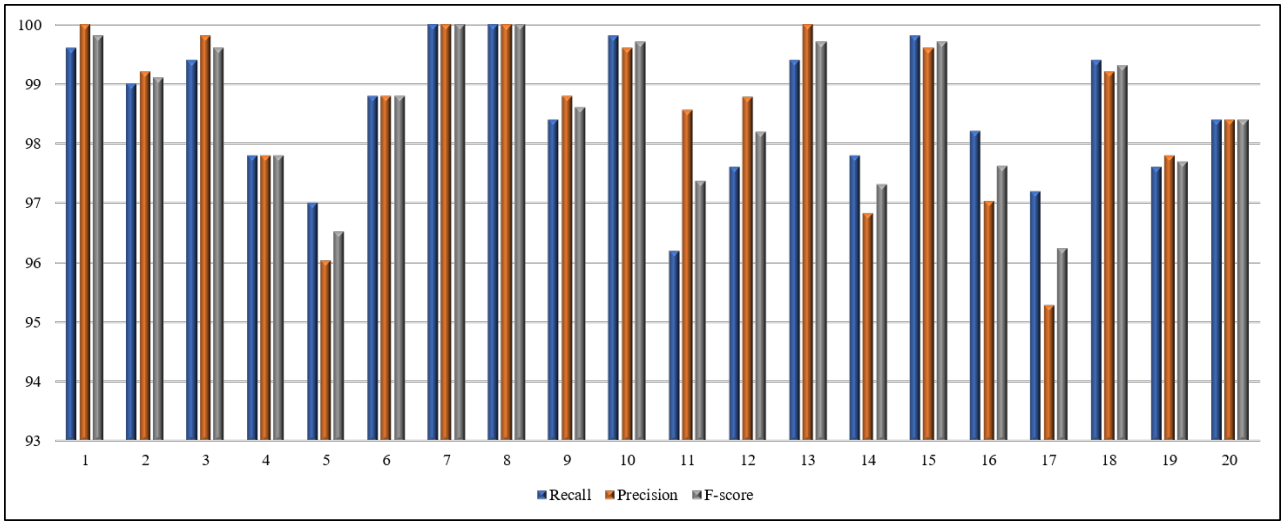

Fig. 3. Result of Class wise recall, precision and F-score on GHIM10K database using VGG19 network architecture.

Fig. 4. Sample images from GHIM10K database. One image from each class.

TABLE I
COMPARISON BETWEEN CNN ARCHITECTURES USING AVERAGE RECALL, PRECISION AND F-SCORE ON GHIM10K DATABASE

Method	*Recall*	*Precision*	*F-score*
AlexNet	96.88	96.56	96.72
VGG16	98.57	98.23	98.40
VGG19	99.38	99.23	99.30

TABLE II
COMPARISON BETWEEN CNN ARCHITECTURES USING AVERAGE RECALL, PRECISION AND F-SCORE ON CALTECH256 DATABASE

Method	*Recall*	*Precision*	*F-score*
AlexNet	87.08	87.31	87.09
VGG16	88.04	88.24	88.03
VGG19	88.63	88.88	88.65

and F-score} over GHIM10K database using VGG19 architecture. Further, to analyse the robustness of CNN features, we employed support vector machine for image classification. Also, we have compared performance of SVM with VGG19. Along with VGG19, we have analysed performance of AlexNet and VGG16 on GBHIM10K database. Fig. 5 shows the comparison between the three CNN architectures and hybrid approach (SVM) over GHIM10K and CalTech256 database. Table I shows the comparison between CNN architectures using average recall, precision and F-score on GHIM10K database. Table I witnessed to the improvement in the accuracy due to the VGG19 architecture.

B. Experiment #2

This experiment comprises use of CalTech256 database for performance evaluation of VGG19 architecture for image classification task. CalTech256 consists of 256 categories, each class is having minimum 80 images. Due to the space limit, it is not possible to show the class wise accuracy, as there are 256 different categories. Instead we discussed average { recall, precision and F-score} over CalTech256 dataset. Further, to analyse the robustness of CNN features, we employed support vector machine for image classification. Also, we have compared performance of SVM with VGG19. Along with VGG19, we have analysed performance of AlexNet and VGG16 on CalTech256 database. Table II shows the comparison between CNN architectures using average recall, precision and F-score on CalTech256 database. From Table II, it can be observed that VGG19 improves the system accuracy. Fig. 5 shows the overall comparison between the three CNN architectures over GHIM10K and CalTech256 datasets.

V. CONCLUSION

In this paper, we fine-tuned the network (AlexNet, VGG16 and VGG19) weight parameters over two state-of-the-art databases namely: GHIM10K and CalTech256 for image clas-

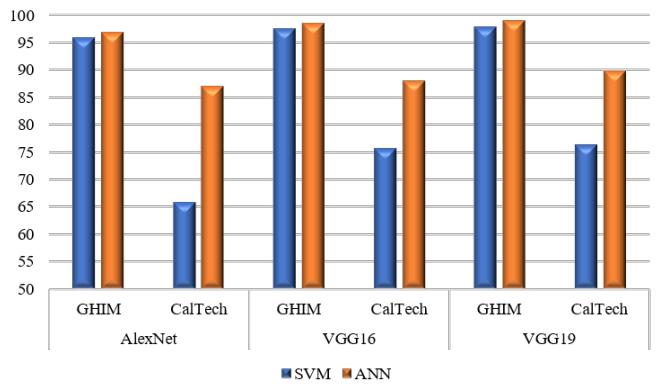

Fig. 5. Overall comparison between the three CNN architectures and hybrid approach (SVM) over GHIM10K and CalTech256 database.

sification task. We compare the performance of these network architectures using three parameters recall, precision and F-score. Further, to analyse the robustness of CNN features, we employed support vector machine for image classification. We have compared performance of SVM with discussed CNN networks. Performance analysis witnessed to the % improvement in the average recall, precision and F-score on both the databases using VGG19 CNN architecture. In future, these fine-tuned network architectures can be used into high level tasks such as object detection, human action recognition etc.

REFERENCES

[1] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov 1973.
[2] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, May 1979.
[3] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, no. 8, pp. 837–842, 1996.
[4] J. Han and K.-K. Ma, "Rotation-invariant and scale-invariant gabor features for texture image retrieval," *Image and vision computing*, vol. 25, no. 9, pp. 1474–1481, 2007.
[5] S. N. Talbar, R. S. Holambe, and T. R. Sontakke, "Supervised texture classification using wavelet transform," in *Signal Processing Proceedings, 1998. ICSP '98. 1998 Fourth International Conference on*, vol. 2, 1998, pp. 1177–1180 vol.2.
[6] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
[7] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1635–1650, 2010.
[8] S. Murala and Q. J. Wu, "Local mesh patterns versus local binary patterns: biomedical image indexing and retrieval," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 929–938, 2014.
[9] S. Murala, R. Maheshwari, and R. Balasubramanian, "Directional binary wavelet patterns for biomedical image indexing and retrieval," *Journal of Medical Systems*, vol. 36, no. 5, pp. 2865–2879, 2012.
[10] S. Murala and Q. J. Wu, "Local ternary co-occurrence patterns: a new feature descriptor for mri and ct image retrieval," *Neurocomputing*, vol. 119, pp. 399–412, 2013.
[11] M. Subrahmanyam and Q. J. Wu, "Spherical symmetric 3d local ternary patterns for natural, texture and biomedical image indexing and retrieval," *Neurocomputing*, vol. 149, pp. 1502–1514, 2015.
[12] S. Murala and Q. J. Wu, "Mri and ct image indexing and retrieval using local mesh peak valley edge patterns," *Signal Processing: Image Communication*, vol. 29, no. 3, pp. 400–409, 2014.
[13] S. Murala and Q. Wu, "Peak valley edge patterns: a new descriptor for biomedical image indexing and retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 444–449.
[14] S. Murala, R. Maheshwari, and R. Balasubramanian, "Local tetra patterns: a new feature descriptor for content-based image retrieval," *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2874–2886, 2012.
[15] M. Subrahmanyam, R. Maheshwari, and R. Balasubramanian, "Local maximum edge binary patterns: a new descriptor for image retrieval and object tracking," *Signal Processing*, vol. 92, no. 6, pp. 1467–1479, 2012.
[16] S. Murala and Q. J. Wu, "Expert content-based image retrieval system using robust local patterns," *Journal of Visual Communication and Image Representation*, vol. 25, no. 6, pp. 1324–1334, 2014.
[17] A. Dudhane, G. Shingadkar, P. Sanghavi, B. Jankharia, and S. Talbar, "Interstitial lung disease classification using feed forward neural networks," in *Advances in Intelligent Systems Research, ICCASP*, vol. 137, 2017, pp. 515–521.
[18] M. Sadafale and S. V. Bonde, "Spatio-frequency local descriptor for content based image retrieval," in *2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, Aug 2017, pp. 1–5.
[19] C. Harris and M. Stephens, "A combined corner and edge detector." 1988.
[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
[21] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.
[22] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Computer Vision – ECCV 2006*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 490–503.
[23] A. Verma, S. Banerji, and C. Liu, "A new color sift descriptor and methods for image category classification," in *International Congress on Computer Applications and Computational Science*, 2010, pp. 4–6.
[24] M. Brown and S. Ssstrunk, "Multi-spectral sift for scene category recognition," in *CVPR 2011*, June 2011, pp. 177–184.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[26] J. D. A. Berg and L. Fei-Fei, "Large scale visual recognition challenge 2010," http://image-net.org/download, 2010, [Online; accessed 29-Jan-2018].
[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[28] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.
[29] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 9, pp. 1075–1088, 2003.