Content-Based Image Retrieval Established on Deep and Handcrafted Features

Faiza Imdad[*], Sana Ul Haq[†], Imtiaz Rasool[‡], Muhammad Wasimuddin[§], Mohammad Omer Farooq[**]

Abstract

The role of image features is crucial in any system. There is a need to enhance Content-Based Image Retrieval (CBIR) systems, but challenges exist in accurately classifying, retrieving, and browsing or mining images. These challenges can be effectively addressed through the extraction of meaningful visual features. Various handcrafted and Deep Learning (DL) techniques have been developed for this purpose, but there has been limited exploration of combining the two approaches. The proposed technique is based on joint use of handcrafted features, i.e., Histogram of Oriented Gradients (HOG), Speed-Up Robust Features (SURF), Bag of Features (BoF), and Local Binary Pattern (LBP), and deep features extracted through AlexNet plus Spatial Pyramid Pooling (SPP) model. The Support Vector Machine (SVM) classifier was used for the classification. The algorithm was evaluated using the Caltech-256 RGB image dataset, achieving an average accuracy of 86.8%. The outcomes demonstrated the benefits of combining handcrafted and DL features, leading to improved accuracy in specific CBIR scenarios.

Keywords: Image Retrieval; Deep Features; Handcrafted Features; Combined Features; Classification.

Introduction

Machine learning represents significant importance within classification. Various techniques leverage deep learning features in image retrieval whereas others utilize content-based aspects like shape, color, and texture. Notable researchers such as Latif et al. (2019), Pathak & Raju (2022), Xie et al. (2015), Hameed et al. (2021), and Kulkarni & Manu (2022) have significant contribution to both Content-Based Image Retrieval (CBIR) and image classification. Within CBIR, Remote Sensing Image Retrieval (RSIR) stands out as significant research area. The

[*] Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, faiza.imdad92@gmail.com

[†] Corresponding Author: Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, sanaulhaq@uop.edu.pk

[‡] Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, imtiazrasoolkhan@uop.edu.pk

[§] Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, wasimuddin@uop.edu.pk

[**] Department of Electronics, University of Peshawar, Peshawar 25120, Pakistan, omer908@uop.edu.pk

PatternNet dataset (Zhou et al., 2018) was tailored for RSIR, consists of 38 classes including 800 images per class. This dataset serves as a benchmark for evaluating both deep learning and traditional methods. CBIR faces challenges of the intention and semantic gaps (Zhou et al., 2017). Sarwar et al. (2019) proposed a novel approach using Bag-of-Words (BoW), where features of visual words are combined from Local Intensity Order Pattern (LIOP) and Local Binary Pattern Variance (LBPV), aiming to enhance the CBIR performance and minimize the semantic gap. The demonstrated results show superiority of this novel method over traditional handcrafted techniques. The design of CBIR system extends to texture image retrieval, where (Pham, 2018) presents a method leveraging feature extraction based on multiscale local extrema and covariance embedding. Utilizing the handcrafted descriptor, Simple Local Extrema Descriptor (SLED), yielded notable performance in retrieval compared to other contemporary techniques. Moreover, CBIR finds application in medical image analysis. Lacoste et al. (2007) proposed the Unified Medical Language System (UMLS) concept to facilitate automatic categorization and extraction of numerous visual instances. Their approach comprised global indexing for image modality access and local indexing for semantic local feature retrieval. Additionally, two fusion strategies were devised. Firstly, enhanced results were achieved for both text and images by developing a basic combination of retrieval for textural and visuals. Secondly, design of a visual modality filter eliminated visually diverged images based on query modality concept. This approach showcased promising outcomes on the Image CLEFmed database (Clough et al., 2006). Extraction of semantic from images based on t-SNE method was proposed by Taheri et al. (2023).

Deep learning techniques have become prominent in CBIR systems (Wan et al., 2014). Razavian et al. (2016) introduced a streamlined pipeline for visual instance retrieval leveraging Convolutional Neural Networks (CNNs), surpassing previous state-of-the-art methods. CNN activations from top layers of large networks serve as high-level descriptors for image content (Babenko et al., 2014). Utilizing these neural codes in image retrieval applications yielded superior results, even on unrelated classification tasks like the ImageNet dataset. Mao et al. (2014) presented a multimodal Recurrent Neural Network (m-RNN) for CBIR, that classify sentences using deep RNN and images using deep CNN, achieving superior performance over other generative methods on standard benchmark datasets. Hashing methods are also popular in CBIR systems with deep learning. Zhang et al. (2015) introduced raw images-based generation of compact and scalable hashing codes using a supervised learning framework. Qayyum et al. (2017) introduced a CNN based Content-Based Medical Image Retrieval (CBMIR) system, surpassing other state-of-the-art methods in classification accuracy. Pre-trained CNN on ResNet-18 was used to develop CBIR system to extract features by Ahmed (2021). However, deep learning models may not consistently improve image retrieval due to issues like noisy training data and suboptimal architectures. Proposed solutions include leveraging large-scale noisy datasets with automatic cleaning methods, utilizing an R-MAC descriptor for deep and differentiable architectures, and training networks via Siamese architectures with triplet loss, resulting in a global image representation ideal for retrieval with promising results.

Some studies have explored the combined use of handcrafted and deep features for CBIR tasks. Zhang et al. (2017) introduced a Combined Deep Handcrafted Visual Feature (CDHVF) based algorithm that presents a unified approach by combining fine-tuned and pre-trained deep CNN models with hand crafted descriptors. Evaluating the CDHVF algorithm on the image CLEF 2016 subfigure classification dataset yielded promising results. In the medical imaging domain, another area of interest is Chest Radiograph Image Retrieval (CRIR) systems. Researchers have employed handcrafted features such as dense SIFT-BoVW, LBP, and Binary, alongside deep features, creating a comprehensive algorithm tested with 443 X-ray query images (Anavi et al., 2015). This combined technique demonstrated promising outcomes.

Previous studies indicate a limited exploration of combining handcrafted and deep features for CBIR tasks, a detailed survey of these methods are listed by Dubey (2021) and Hameed et al. (2021). The proposed technique integrates handcrafted features from HOG, SURF, BoF, and LBP, with deep features extracted via the AlexNet with Spatial Pyramid Pooling (SSP) model. Classification was performed using an SVM classifier. Testing was conducted on the Caltech-256 dataset, comprising 257 classes. The proposed framework yielded superior results for CBIR tasks by encompassing both low- and high-level visual content information. The subsequent sections detail the methodology, experimental findings, and conclusions.

Methodology

A CBIR system is proposed based on deep features and handcrafted descriptors. This system queries an image and performs classification utilizing visual instances. The function of this system is shown in Figure 1. The deep features were extracted using AlexNet model combined with SPP layer. In addition, the handcrafted features were also extracted. The two sets of features were then combined and SVM classifier was used for classification. The Caltech-256 dataset was used for the analysis.

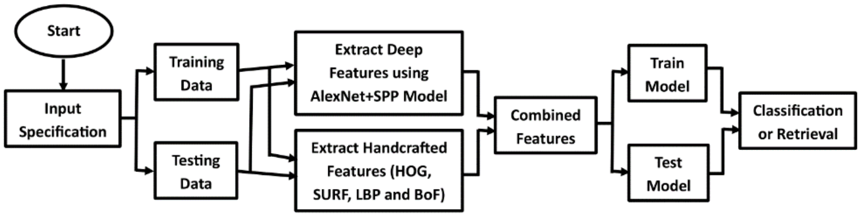
Start
Input Specification
Training Data
Testing Data
Extract Deep Features using AlexNet+SPP Model
Extract Handcrafted Features (HOG, SURF, LBP and BoF)
Combined Features
Train Model
Test Model
Classification or Retrieval

Figure 1: Flow diagram of the proposed CBIR system

Caltech-256 Dataset

Numerous researches have been conducted in the last decade on object recognition (Sukanya et al., 2016). Several databases have been recorded for this purpose including the Coil (Everingham et al., 2015), MIT-CSAIL (Opelt & Pinz, 2005), PASCAL VOC (Russell et al., 2008), Caltech-6 and Caltech-101 (Lowe, 2004). Griffin et al. (2022) introduced the Caltech-256 dataset after Caltech-101 (Li et al., 2022). The Caltech-256 was used in this research which consists of 257 classes (256 object classes and a clutter class). The Caltech-256 has an average number of 119 while a minimum number of 80 samples per class. In comparison, the Caltech-101 has 102 categories, and has an average number of 90 samples while minimum number of 31 samples per class. The Caltech-256 dataset has variety of images with diversity in lighting condition, poses, backgrounds, image sizes and camera taxonomy. Images in this dataset can be used as provided and does not require further editing or preprocessing. This makes Caltech-256 database better compared to Caltech-101 dataset.

Deep Features

Krizhevsky et al. (2012) presents Alexnet, a CNN model, as shown in Figure 2 with butterfly image, taken from Caltech-256 dataset. The model is designed with five convolutional layers consisting of size 11x11, three layers of max pooling, connected to three fully connected layers. ReLU activation function is used in both the convolutional and fully connected layers. Filters uses different formulations and are also shown in the figure. Finally, SoftMax is used to classify the sample based on its features.

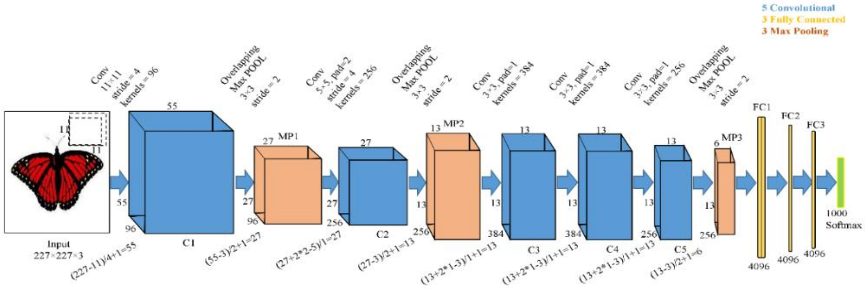

Figure 2: AlexNet model architecture

SPP Layer

The spatial pyramid pooling layer makes bins of the input sample, as shown in Figure 3. These bins hold the maximum information that can help in accurate query image classification.

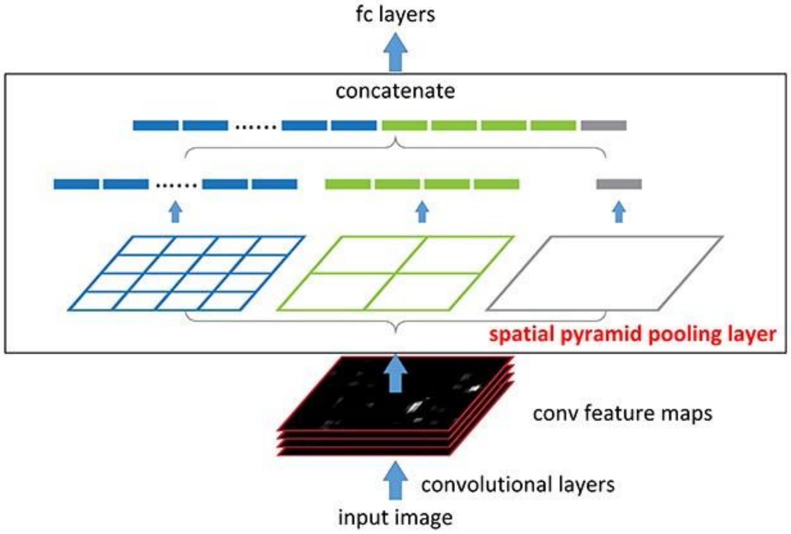

Figure 3: SPP model architecture

Proposed Model

The proposed model consists of AlexNet model combined with SPP layer, as shown in Figure 4. It consists of five convolutional layers, three max pooling layers, an SPP layer, and three fully connected layers before the SoftMax is applied, that extracts deep features.

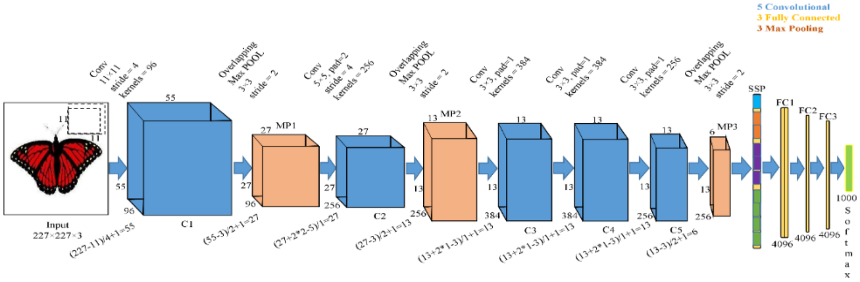

Figure 4: AlexNet model combined with SPP layer.

Handcrafted Descriptors

The proposed method uses HOG, SURF, BoF and LBP handcrafted descriptors. Number of gradient orientation occurrences in local portion of the image are counted by the HOG features descriptor. Another computer vision tool to detect local feature is SURF, an enhanced version of Scale-Invariant Feature Transform (SIFT). Image classification is achieved by the LBP, another local visual feature descriptor. Caltech-256 database was used by these descriptors and obtained the local handcrafted features.

Combined Deep and Handcrafted Features

These feature descriptors were merged to achieve a proficient feature set. After extracting the two sets of features, the classification was performed using the SVM classifier (Platt, 1999).

Experimental Results

In the first phase, handcrafted features from HOG, SURF, BoF and LBP are used to perform the classification. In the second phase, the classification was performed using deep features extracted by different deep models, i.e., Caffe-ref, VGG-f, and VGG-19, AlexNet and AlexNet+SPP. The third phase integrates the deep features CNN and individual handcrafted features HOG, SURF, BoF, and LBP resulting in CNN+HOG, CNN+SURF, CNN+BoF and CNN+LBP. The last phase combines these deep and handcrafted features into CNN+HOG+SURF+BoF+LBP. These experiments were performed with the SVM classifier for classification.

The pretrained AlexNet model using ImageNet dataset was used for classification on the Caltech-256 dataset, where the ratio of training to testing images was 70% to 30%. The ReLU activation function used for faster training, and Stochastic Gradient Descent (SGD) was used an optimization algorithm for CNN model presented in this study. The network was trained with batch size 128, momentum value of 0.9, and 0.01

learning rate. The dropout regularization, set to 0.5, was applied to the two fully connected layers.

Performance on Handcrafted Features

Table 1 shows the performance metrics and their values achieved by the proposed model for different handcrafted features. The values of specificity, sensitivity, accuracy, and Matthew's Correlation Coefficient (MCC) (Chicco et al., 2021) are based on the Caltech-256 dataset. The accuracy and specificity outperform by HOG features, while SURF features outperformed others in terms of sensitivity and MCC.

Table 1: Performance metrics of different handcrafted features.

Handcrafted Features	Specificity (%)	Sensitivity (%)	Accuracy (%)	MCC
HOG	69.9	73.0	72.0	0.57
SURF	69.1	75.1	70.3	0.59
BoF	41.9	71.5	71.0	0.29
LBP	65.5	73.6	71.5	0.53

Performance on Deep Features

The performance metrics achieved by the proposed model for deep features are given in Table 2. The deep features were extracted using different deep learning models, i.e., Caffe-ref, VGG-f, VGG-19, AlexNet and AlexNet with SPP layer (AlexNet+SPP). The deep features obtained through AlexNet+SPP model outperformed all other deep features in terms of different performance metrics. It was followed by deep features extracted through AlexNet model.

Table 2: Performance metrics of different deep features.

Deep Features Extraction Model	Specificity (%)	Sensitivity (%)	Accuracy (%)	MCC
Caffe-ref	73.6	78.9	79.9	0.67
VGG-f	74.7	77.5	79.8	0.68
VGG-19	73.0	75.5	80.2	0.69
AlexNet	78.9	80.0	80.3	0.70
AlexNet+SPP	79.9	81.8	80.6	0.80

Performance on Combined Deep and Handcrafted Features

In the next step, the deep features extracted through AlexNet and AlexNet+SPP models were combined with handcrafted features to enhance classification results given in Table 3. Combining deep and handcrafted feature improved the overall performance. Better accuracy, sensitivity and MCC results were achieved using the AlexNet+SPP and

HOG features. In terms of specificity, the AlexNet and HOG features provided better results. In general, the AlexNet+SPP deep features provided better results as compared to AlexNet combined deep and individual handcrafted features. At final stage, the deep features extracted by AlexNet+SPP model were combined with all the handcrafted features. Classification accuracy of 86.8% achieved by the proposed model using the Caltech-256 dataset.

Table 3: Performance metrics of different deep plus handcrafted features.

Deep + Handcrafted Features	Specificity (%)	Sensitivity (%)	Accuracy (%)	MCC
AlexNet+HOG	80.5	81.7	82.5	0.77
AlexNet+SURF	79.6	81.0	81.6	0.75
AlexNet+BoF	79.8	80.0	80.6	0.66
AlexNet+LBP	78.4	80.6	80.1	0.65
AlexNet+SPP+HOG	80.0	82.8	83.9	0.80
AlexNet+SPP+SURF	80.1	82.6	82.5	0.75
AlexNet+SPP+BoF	79.5	81.1	82.0	0.69
AlexNet+SPP+LBP	79.6	81.0	80.7	0.66
AlexNet+SPP+HOG+ SURF+BoF+LBP	82.8	85.7	86.8	0.90

Table 4: Classification performance of different handcrafted, deep, and combined features.

Method	Features	Classification Accuracy (%)
Handcrafted	HOG	72.0
	SURF	70.3
	BoF	71.0
	LBP	71.5
Deep	Caffe-ref	79.9
	VGG-f	79.8
	VGG-19	80.2
	AlexNet	80.3
	AlexNet+SPP	80.6
Deep + Handcrafted	AlexNet+HOG	82.5
	AlexNet+SURF	81.6
	AlexNet+BoF	80.6
	AlexNet+LBP	80.1
	AlexNet+SPP+HOG	83.9
	AlexNet+SPP+SURF	82.5
	AlexNet+SPP+BoF	82.0
	AlexNet+SPP+LBP	80.7
Proposed	AlexNet+SPP+HOG+ SURF+BoF+LBP	86.8

The classification accuracies obtained for the deep and handcrafted features, and their different combinations are summarized in Table 4. Better classification performance is observed as compared to the handcrafted features. The combination of deep and handcrafted features further improved the classification performance, where best classification result was achieved with the deep features and handcrafted features combination.

Conclusion

In this study, a combined deep and handcrafted features-based classification technique is proposed with SVM classifier. Multiple experiments were performed on Caltech-257 dataset to assess the performance of the proposed model. In the first experiment, among handcrafted features method, i.e., SURF, HOG, BoF and LBP, the HOG features performed better in terms of accuracy and specificity, while SURF features outperformed others in terms of sensitivity and MCC. In the second experiment, deep features extracted through different deep CNN models, i.e., Caffe-ref, VGG-f, VGG-19, AlexNet and AlexNet+SPP were used for classification. The deep features obtained through AlexNet+SPP model outperformed all other deep features in terms of different performance metrics. It was followed by deep features extracted through AlexNet model.

In third experiment, handcrafted features were combined with deep features extracted through AlexNet and AlexNet+SPP were combined with handcrafted features. Initially, the deep features were extracted through AlexNet and AlexNet+SPP models and were combined with individual handcrafted features. Furthermore, deep features extracted through AlexNet+SPP model were combined with four handcrafted features, i.e., SURF, HOG, BoF and LBP yielding improved overall performance when deep and handcrafted features were combined. The best accuracy, sensitivity and MCC results were achieved using the AlexNet+SPP and HOG features. In terms of specificity, the AlexNet and HOG features provided better results. In general, the AlexNet +SPP deep features provided better results as compared to AlexNet deep features when combined with individual handcrafted features. The proposed approach provided the best classification accuracy of 86.8% when deep features extracted through AlexNet+SPP model were combined with four handcrafted features. In future various pre-trained CNN models, e.g., MobileNet, ResNet-50, Inception and Xception, can be used to extract the visual features to further improve the performance of the proposed model with fine tuning the model’s parameters.

References

Ahmed, A. (2021). Pre-trained CNNs models for content based image retrieval. *International Journal of Advanced Computer Science and Applications, 12*(7), 200-206.

Anavi, Y., Kogan, I., Gelbart, E., Geva, O., & Greenspan, H. (2015). A comparative study for chest radiograph image retrieval using binary texture and deep learning classification. *Proceedings of international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 2940-2943). IEEE.

Babenko, A., & Lempitsky, V. (2015). Aggregating deep convolutional features for image retrieval. *arXiv preprint arXiv:1510.07493*.

Babenko, A., Slesarev, A., Chigorin, A., & Lempitsky, V. (2014). Neural codes for image retrieval. *Proceedings of European Conference on Computer Vision* (pp. 584-599). Springer.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE Access, 9*(1), 78368-78381.

Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T. M., Jensen, J., et al. (2006). The CLEF 2005 Cross–Language Image Retrieval Track. *Proceedings of Workshop of the Cross-Language Evaluation Forum for European Languages* (pp. 535-557). Vienna: Springer Berlin Heidelberg.

Dubey, S. R. (2021). A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(5), 2687--2704.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International journal of computer vision, 111*, 98-136.

Griffin, G., Holub, A., & Perona, P. (2022). *Caltech 256 (1.0).* CaltechDATA.

Hameed, I. M., Abdulhussain, S. H., & Mahmmod, B. M. (2021). Content-based image retrieval: A review of recent trends. *Cogent Engineering, 8*(1), 1927469.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Communications of ACM, 60,* 84-90.

Kulkarni, S., & Manu, T. M. (2022). Content based image retrieval: A literature review. *International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1580-1587). IEEE.

Lacoste, C., Lim, J.-H., Chevallet, J.-P., & Le, D. H. (2007). Medical-image retrieval based on knowledge-assisted text and image indexing. *IEEE Transactions on Circuits and Systems for Video Technology, 17*(7), 889-900.

Latif, A., Rasheed, A., Sajid, U., Ahmed, J., Ali, N., Ratyal, N. I., et al. (2019). Content-based image retrieval and feature extraction: a comprehensive review. *Mathematical Problems in Engineering, 2019*, 1-21.

Li, F.-F., Andreeto, M., Ranzato, M., & Perona, P. (2022). *Caltech 101 (1.0)*. CaltechDATA.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*, 91-110.

Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2014). Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

Opelt, A., & Pinz, A. (2005). Object localization with boosting and weak supervision for generic object. *Proceedings of Scandinavian Conference on Image Analysis* (pp. 862-871). Springer.

Pathak, D., & Raju, U. S. (2022). Content-based image retrieval for super-resolutioned images using feature fusion: Deep learning and hand crafted. *Concurrency and Computation: Practice and Experience, 34*(22), e6851.

Pham, M.-T. (2018). Efficient texture retrieval using multiscale local extrema descriptors and covariance embedding. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 564-579.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods,* 185-208.

Qayyum, A., Anwar, S. M., Awais, M., & Majid, M. (2017). Medical image retrieval using deep convolutional neural network. *Neurocomputing, 266*, 8-20.

Razavian, A. S., Sullivan, J., Carlsson, S., & Maki, A. (2016). Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications, 4*(3), 251-258.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International journal of computer vision, 77*, 157-173.

Sarwar, A., Mehmood, Z., Saba, T., Qazi, K. A., Adnan, A., & Jamal, H. (2019). A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine. *Journal of Information Science, 45*(1), 117-135.

Sukanya, C. M., Gokul, R., & Paul, V. (2016). A survey on object recognition methods. *International Journal of Science, Engineering and Computer Technology*, *6*(1), 48-52.

Taheri, F., Rahbar, K., & Beheshtifard, Z. (2023). Content-based image retrieval using handcraft feature fusion in semantic pyramid. *International Journal of Multimedia Information Retrieval*, *12*(2), 21.

Wan, J., Wang, D., Hoi, S. C., Wu, P., Zhu, J., Zhang, Y., et al. (2014). Deep learning for content-based image retrieval: A comprehensive study. *Proceedings of the ACM international conference on Multimedia*, (pp. 157-166).

Xie, L., Hong, R., Zhang, B., & Tian, Q. (2015). Image classification and retrieval are one. *Proceedings of ACM International Conference on Multimedia Retrieval*, (pp. 3-10).

Zhang, J., Xia, Y., Xie, Y., Fulham, M., & Feng, D. D. (2017). Classification of medical images in the biomedical literature by jointly using deep and handcrafted visual features. *IEEE journal of biomedical and health informatics*, *22*(5), 1521-1530.

Zhang, R., Lin, L., Zhang, R., Zuo, W., & Zhang, L. (2015). Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Transactions on Image Processing, 24*(12), 4766-4779.

Zhou, W., Li, H., & Tian, Q. (2017). Recent advance in content-based image retrieval: A literature survey. *arXiv preprint arXiv:1706.06064*.

Zhou, W., Newsam, S., Li, C., & Shao, Z. (2018). PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing, 145*, 197-209.