Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2022.018822
Article

Tech Science Press

Enhancing Scalability of Image Retrieval Using Visual Fusion of Feature Descriptors

S. Balammal@Geetha*, R. Muthukkumar and V. Seenivasagam

Department of Information Technology, National Engineering College, Kovilpatti, 628 503, Tamil Nadu, India
*Corresponding Author: S. Balammal@Geetha. Email: geethabalammalmanian@gmail.com
Received: 22 March 2021; Accepted: 10 May 2021

Abstract: Content-Based Image Retrieval (CBIR) is an approach of retrieving similar images from a large image database. Recently CBIR poses new challenges in semantic categorization of the images. Different feature extraction technique have been proposed to overcome the semantic breach problems, however these methods suffer from several shortcomings. This paper contributes an image retrieval system to extract the local features based on the fusion of scale-invariant feature transform (SIFT) and KAZE. The strength of local feature descriptor SIFT complements global feature descriptor KAZE. SIFT concentrates on the complete region of an image using high fine points of features and KAZE ponders on details of a boundary. The fusion of local feature descriptor and global feature descriptor boost the retrieval of images having diverse semantic classification and also helps in achieving the better results in large scale retrieval. To enhance the scalability of image retrieval bag of visual words (BoVW) is mainly used. The fusion of local and global feature representations are selected for image retrieval for the reason that SIFT effectively captures shape and texture and robust towards the change in scale and rotation, while KAZE have strong response towards boundary and changes in illumination. Experiments conducted on two image collections, namely, Caltech-256 and Corel 10k demonstrate the proposed scheme appreciably enhanced the performance of the CBIR compared to state-of-the-art image retrieval techniques.

Keywords: Content-based image retrieval; scale-invariant feature transform; bag of visual word; scalability; relevance feedback

1 Introduction

The progressive advancement of technology has led to a rapid increase in the collection of digital images as well as the image repository. Retrieval of images according to the user's objective from the haystack of a large database is a tedious process. CBIR [1] provides a hopeful solution to retrieve the desired image from large-scale databases. Effective feature extraction is the fundamental need of the retrieval process. All CBIR applications need promising extracted features since their discriminative power improves retrieval performance.

Generally, the features are classified as global features and local features. Global features describe the feature distribution of an entire image. It provides global descriptions based on color, shape, and texture. Several studies have been done on feature extraction based on color [2,3], shape [4,5], and texture [6,7]. Local features concentrate on salient image patches or keypoints and are powerful, which allows them to achieve higher rates of image retrieval even in the case of clutter and occlusion. Several approaches such as Scale Invariant Feature Transform SIFT [8], Speeded-Up Robust Features SURF [9], and Oriented FAST and rotated BRIEF ORB [10] analyses the local features.

It has been observed that a single feature-based representation is not suitable for attaining a higher rate of retrieval performance. Hence several works are done based on a combination of multiple features. Some research works combined color with other features [11–14]. Numerous approaches have investigated the comparison of local features and global features and proved that the fusion of local features results in improved retrieval efficiency and immense discriminative power compared to the fusion of global features [15–17].

However, the major constraint of global features is the segmentation for the general description of an image. Further, it is more likely to fail in the case of occlusion and clutter. It is sensitive to locality and fails to recognize vital visual characteristic which makes global features too stringent for image representation [18,19]. In addition to this, the global method does not concentrate on the details of the shape. It considers only the main properties of objects. However, the local descriptors give more emphasis on the details [20].

The aforementioned facts motivated us to explore the fusion of local features. It is studied that the local features are capable of capturing minuscule features in images moreover, they are compact and expressive. Local feature determines salient keypoints in an image and provides more robustness to occlusion and clutter. To the authors' knowledge, the proposed work is the first to perform the feature fusion of SIFT and KAZE for image retrieval.

Another major hindrance in image retrieval is bridging the semantic gap. To bridge the semantic gap relevance feedback is used. In large image datasets, the number of local features extracted for every image may be enormous. To solve this problem, BoVW [21,22] is proposed. It quantizes the descriptor and generates visual words.

The key contribution of this paper is summarized as:

- The fusion of SIFT and KAZE descriptor preserves the unique property of an image representation.
- The introduction of the BoVW framework enhances scalability.
- The inclusion of an RF system based on the user's feedback reduces the semantic gap.

The rest of this paper is organized as follows. Section 2 presents the related works. The general idea of the BoVW model, k-means, and SVM is discussed in Section 3. Section 4 addresses the local feature descriptors. Section 5 presents the proposed method. Experimental results are discussed in Section 6. Section 7 addresses the conclusion and future work.

2 Related Work

CBIR has emerged extensively since the 1990s [23] and still triggers the attention of researchers towards scalable image retrieval. The pioneering works on SIFT [8] and BoVW [24] paved the way for a significant advance in CBIR. In this section, we briefly describe the most widely used, state-of-the-art methods for CBIR. Hiremath and Pujari [25] proposed a framework in which color, texture, and shape information are combined to generate a forceful feature set for image retrieval.

Ashraf et al. [26] introduced a technique based upon YCbCr color including a canny edge histogram and discrete wavelet transform. Additionally, the execution of different wavelets is done to find the suitable wavelet. Image retrieval is carried out with the help of an artificial neural network. Ahmed et al. [27] analysed the RGB channel L2 spatial color which fuses color information spatial details with the shape of extracted features and object recognition.

Bu et al. [28] presented a new image retrieval method based on texture features with multi-resolution multi-directional (MRMD) filter to separate the low and high- frequency information. Karakasis et al. [29] proposed a CBIR technique that employs an affine moment to the image invariants that lies in the local areas. Mehmood et al. [30] presented a technique in which local and global histograms of visual words are combined. The local rectangular region is used to build a histogram which provides a feasible way to capture the semantic details. Mehmood et al. [31] proposed a technique in which the weighted average of triangular histograms of visual words is computed. The image spatial contents are added to the inverted index of the BoVW model. Dimitrovski et al. [32] presented a technique in which the visual Dictionary is constructed by predictive clustering trees (PCTs). They used random forests of predictive clustering trees to increase stability. The Dictionary and indexing structure are characterized by PCT's. Swathi Rao [33] presented a technique in which image feature vectors are calculated using different classes of dense SIFT and they are quantized to visual words. Liu et al. [34] analysed the associations among the vocabulary size, classification performance, and universality of codebooks. Subhransu Maji et al. [35] presented a technique with nonlinear kernel SVM which show significant progress in image classification. Mohammed Alkhawlani et al. [36] presented a retrieval system that uses the local feature descriptors SIFT and SURF along with BoVW for efficient image retrieval. The image signature produced by the system is invariant to rotation and scale. The system utilizes k-means as a clustering algorithm and includes SVM for the retrieval of relevant images. Sharif et al. [37] combined SIFT and BRISK to form a single dictionary and lessen the semantic gap. The fusion of visual words and the suitable feature percentage selection reduces the run time.

3 Bag-of-Visual Word (BoVW) Model

The Bag of visual word model is shown in Fig. 1. This model provides the image structure from low-level features to high-level features. It represents small characteristic regions of complex images without explicitly modeling the object. Visual vocabulary is constructed by the extraction of features in training images. To construct the visual vocabulary, clustering is integrated. Every feature in a cluster is termed a visual word. The feature detection is done in the given training image and the same is assigned to their cluster centers to form the visual vocabulary. The quantized feature represents the normalized histogram and is termed as a vector.

3.1 Building Vocabulary

Feature extraction was performed on all the training images. Visual vocabulary consists of clustered features and these features are vector quantized. The visual word defines every cluster. Vocabulary terms are the codes in the codebook. The count of each term that appears in an image creates a normalized histogram that represents the BoVW.

3.2 Classification and Clustering

In this work, we present a method for producing a robust set of features by the utilization of SIFT and KAZE. K-means clustering computes the nearest neighbor points and the cluster center. It makes use of the approximation of the nearest neighbor method for computation and it scales to a similar large size vocabulary [38,39]. As the SIFT feature is based on the scale-space of Gaussian it gives equal

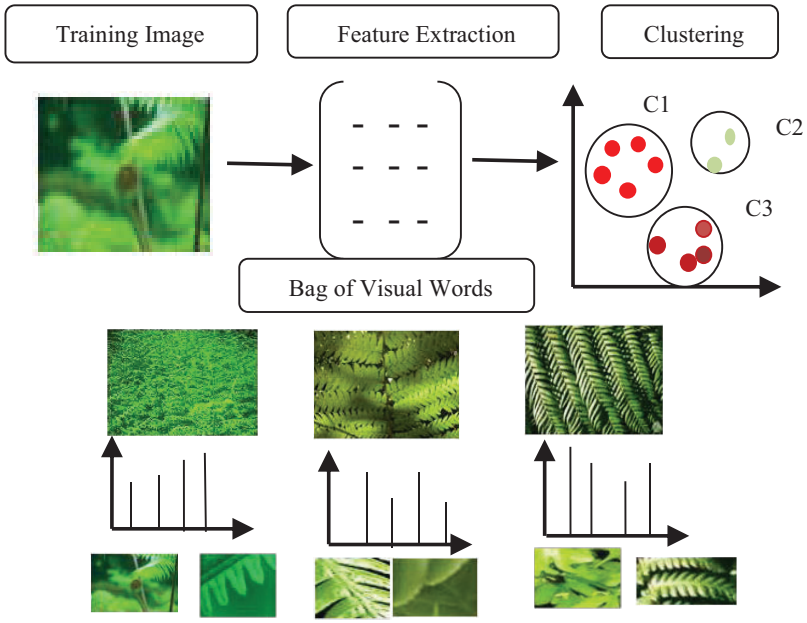

Figure 1: Bag of visual words

significance to the features on the image boundaries and those within it. Whereas the KAZE feature is based on the nonlinear scale-space, it conserves the features present in the boundaries of the image. Thus, the combination of SIFT and KAZE yields a fine merge of features in boundary and look. To perform classification SVM is used. Generally, SVM provides high-performance accuracy in classification when compared with other algorithms [40]. Concerning some datasets, the efficiency of SVM is very sensitive based on how the cost parameter and kernel parameters are set. The kernel trick allows the SVM model to make separations even in the case of very complex boundaries. If the linear separable problem is concerning then the choice of linear SVM is good. Unfortunately, most of the problems are nonlinear in such cases the selection of the cost parameter(C) and kernel parameter is noteworthy.

Here, we choose the Hellinger kernel or Bhattacharya coefficient [41]. This is mentioned in Eq. (1).

$$K(h(i),\ h'(i)) = \sqrt{\sum h(i)h'(i)} \tag{1}$$

where h and h' are normalized histograms of image i.

4 Feature Extraction Using Sift and Kaze

Various applications in image processing emphasize that the obtained feature should possess good uniqueness against diverse image transformations [42–45]. The SIFT and KAZE algorithms are discussed as follows.

4.1 Scale Invariant Feature Transform (SIFT)

SIFT detects the salient feature points and remains invariant to image scale and rotation. It also provides resistance for diverse image transforms including illumination changes, occlusion, and several affine transform. It entails the following four steps.

4.1.1 Determination of Scale-Space Extrema

This generates a multiple-scale pyramid of the original image. This process eliminates the details that are not present at different scales. So, the image is left with information that is invariant to scale. This is achieved by applying Gaussian blur to the image. The algorithm searches for all image regions and scales, which can be invariant to orientation. These are evaluated by applying a Gaussian scale-space kernel, which produces various blurred regions of the original image.

$$G(x,\ y,\ \sigma) = \frac{1}{2\pi\sigma^2}\ \ e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{2}$$

Gaussian convolution and interpolations are performed to create Gaussian scale space. Then images based on difference-of-gaussian (DoG) are computed by approximation of Laplacian of Gaussian and down-sampling.

$$DoG(x,\ y,\ \sigma) = L_1(x,\ y,\ K_1\sigma) - L_2(x,\ y,\ K_2\sigma) \tag{3}$$

To determine the keypoint, the sample point is evaluated to its eight neighbors on the same scale and nine neighbors on the adjoining scale. This shows that at different scales, every pixel will be examined thoroughly and the keypoint is chosen based on the maximum value in that scale of frequency, resulting in invariance in scale.

4.1.2 Determination of Keypoint Localization

The computed numbers of keypoints are increasing, and it is crucial to eliminate the unstable keypoints, which are degraded by noise and have low contrast. Similarly, the detected edge region that is prone to noise should be discarded. This is done by applying the Harris corner detector, and maximum gradients in all scales are detected.

4.1.3 Determination of Orientation Assignment

To achieve strong rotation invariance, each keypoint is assigned an orientation. It considers the magnitude and gradient directions of an image to compute the histogram. The histogram having a maximum peak in the particular direction is investigated and assigned as the orientation for the keypoint.

4.1.4 Keypoint Description

In each region of the keypoint, a neighborhood of 16x16 is chosen. It is separated into a size of 4×4 containing sub-blocks of 16. 8-bin histograms are generated for each sub-block. The histogrammed descriptor vector results in 128 elements. The magnitude calculation is based on the gradient orientations of weights. For matching of keypoint the Euclidean distance between descriptors is calculated. The keypoints matched are called control points, and these are employed to enhance the image transformation.

4.2 KAZE

The KAZE algorithm detects and describes the image feature based on nonlinear diffusion filtering operation and conductivity function. It makes use of nonlinear scale-space. The conductivity function proposed by Perona and Malik captures the gradient of the original image, where the gradient is the smoothed Gaussian function version concerning time. To construct adjustable blur, image features of nonlinear diffusion filtering along with the additive operator Splitting method are used. In SIFT, the Gaussian approach causes blurring and does not preserve the image's natural boundary, but in KAZE, all scale levels of noise are smoothed to the same degree to make blur adaptive to image features.

$$\frac{\partial L}{\partial t} = \mathrm{div}(c(x,\ y,\ t).\nabla L) \tag{4}$$

where div and ∇ are divergence and gradient operators, respectively, c is the conductivity function and t is the scale parameter.

4.2.1 Nonlinear Scale-Space Extrema

KAZE utilizes the combination of nonlinear diffusion filtering with a conductivity function. The conductivity function of Perona and Mallik is represented as:

$$c(x,\ y,\ t) = g(|\nabla L\sigma(x,\ y,\ z)|) \tag{5}$$

where c is the conductivity function, g is the gradient, L is an original image, σ represents the amount of blur and t is time. At edges, the conductivity function helps in the diffusion reduction and achieves further smoothening of regions when compared to edges.

4.2.2 Normalized Keypoint Localization

The determinant of the Hessian (DoH) matrix is scale-normalized and the response is computed by KAZE. DoH is a blob detection method with automatic scale selection. The responses with the maximum value formulate the possible keypoints.

$$L_{Hess} = \sigma^{2(LxxLyy - Lxy^2)} \tag{6}$$

4.2.3 Vector Orientation Assignment

In this process, the dominant orientation is computed in a circular area by a window of sliding orientation having a size of π/3 and a radius of 6 s where s is the scale. The first order derivatives Lx and Ly are computed and weighted with the keypoint of the centered Gaussian. These two responses are summed up by a segment of the sliding circle resulting in the dominant orientation. Then, the longest vector with the dominant orientation is assigned as the orientation of that keypoint.

4.3 KAZE Complements SIFT

SIFT deals with the entire region of the image with the salient keypoint, where KAZE focuses on the details present in the boundary. By this combination of both KAZE and SIFT, the following characteristics are achieved.

- Definite boundary representation.
- Detection of salient keypoint.
- Property of uniqueness.

However, SIFT can smooth information and noises to the same degree, but it fails to preserve the boundary, whereas KAZE preserves the boundary. The illustration of SIFT and KAZE keypoints is shown in the Fig. 2, which depicts that the keypoints of KAZE concentrate on boundary and SIFT looks at all scales for sharp discontinuities, thus capturing salient keypoints in the whole image and maintaining definite boundary representation.

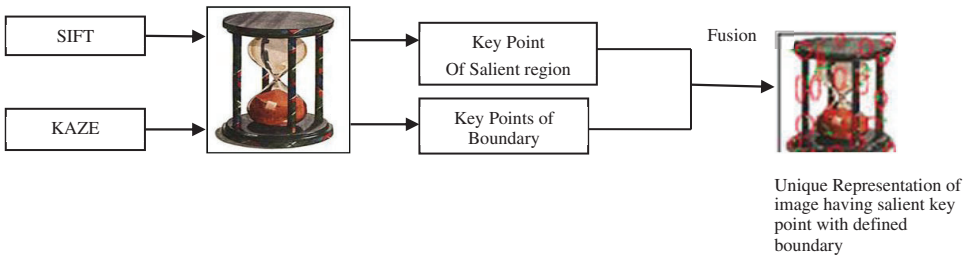

Figure 2: KAZE complementing SIFT

In an image when compared to other areas, the salient region is sparser, and the keypoints present in this region have the capacity of unique identification of an image. SIFT helps to detect these salient keypoints and attain better responses in these regions.

To differentiate images against the background, the keypoints near the object boundaries play a vital role. The utilization of KAZE keypoints addresses this requirement and achieves the property of uniqueness. However, feature descriptor plays a vital role in CBIR semantic gap is the most prominent concern which influence the performance of a CBIR system. To bridge this gap relevance feedback (RF) is introduced with this forceful image representation by fusion of distinctive local and global features from the image. The RF method adds user feedback to enhance retrieval performance and provides related images [46–50]. Hence the RF is integrated with the proposed system, and it performs an iterative process. It refines or modifies the original image interactively to yield more accurate results.

5 Visual Word Fusion of SIFT and KAZE Using BoVW

The method deploys SIFT and KAZE techniques to extract keypoints and calculate feature descriptors. Fig. 3. shows the visual word fusion of SIFT and KAZE using BoVW.

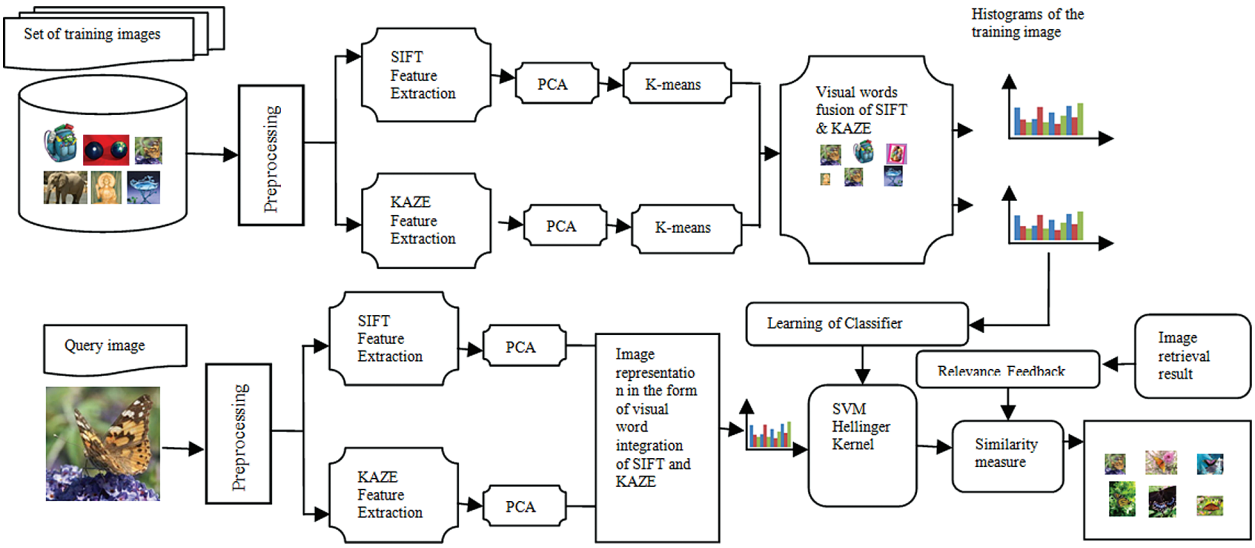

Figure 3: Visual word fusion of sift and kaze

5.1 Preprocessing

The acquired image set needs some preprocessing methods such as converting RGB image to a gray-scale image and image resizing. Grayscale image provides better density details. Digital filtering is done to eliminate noise and inconsistency and normalization is achieved with scaling.

5.2 Compute Visual Vocabulary with SIFT and KAZE

First, the image features are extracted as keypoints by applying SIFT and KAZE feature descriptors. Extracted features contain vast local information and predominant image patches. The spatial structure and the local orientation distribution of surrounded keypoints are captured by SIFT. The multi-neighborhood strategy is applied to find KAZE features. In the case of a large volume of data, the feature extracted may contain large dimension of redundant and correlated data. Hence, it cannot be processed

directly. In this case, we employ principal component analysis (PCA) to reduce the dimensionality of features to achieve better precision. SIFT and KAZE are fused with the help of canonical correlation analysis in a BoVW model, which enhances classification performance.

5.3 BoVW and K-Means Clustering

The BoVW approach creates the vocabulary based on feature descriptors vector quantization. K-means Clustering is the practiced to cluster the descriptors. Then histogram is computed for these clustered descriptors.

- Select initial cluster centroids for vocabulary randomly.
- Calculate the distance between each feature vector and the centroid of the clusters.
- Allocate every feature vector to the cluster with the nearest centroid (minimum distance).
- Recalculate each centroid as the mean of the objects allocated to it.
- Repeat the previous two steps until no change.

As the allocation of the related feature to the same cluster is done, visual words are formed and stored in the codebook. Indexed visual words are formed with the gathering of all similar features within the same code and represent a visual dictionary. In large scale image retrieval systems, BoVW uses index pruning to reduce the retrieval cost. The idea is to recognize and remove the images which are not likely to contribute to top results.

5.4 Image Classification

SVM is then used to classify the images, with the concatenated histograms of code words in *N*dimensional space. Kernel function helps to find the data with no obvious fixed dimensions. The resultant histograms of fused visual words must be normalized and then the SVM Hellinger kernel is applied. It multiplies suitable signatures to obtain any other kernel. It is also known as Bhattacharyya's coefficient and the corresponding metric is the Hellinger distance. The computational cost is low in the SVM Hellinger kernel. It explicitly calculates the feature map, instead of calculating the kernel values, which makes the classifier, to remain linear.

The cross-validation is done on the training dataset and the finest value for the regularization parameter C is chosen as 10-fold. For k number of classes the approach of one-against-one is used to train the data construct $(k-1)/2$ classifiers using two classes. The C parameters control how much misclassification is allowed in the SVM classification.

5.5 Relevance Feedback

The retrieved images are rearranged according to the RF fed back from the user. It represents the progression of refining the results returned by the CBIR system in a given iteration of an interaction session. In the proposed method, the user assigns positive samples to the retrieved relevant images and negative samples to the irrelevant images. Further, based on these positive and negative samples, the proposed method refines the image retrieval results. These steps are repeated in different iterations and user preference is learned according to the positive samples. Thus, the proposed method reduces the semantic gap. The search strategy of image retrieval is improved with the increased number of iterations and the preference of user is studied in this iterations.

5.6 Algorithm

The following algorithm illustrates the efficient visual dictionary of images for different features and proposes a high- performance classification model.

Algorithm:

Input:

1. $m \rightarrow$ Set of large-scale database
2. $n \rightarrow$ Image for classification

a. Output:

3. $y \rightarrow$Set n number of classified images
4. BoVW model ()

Begin

a. Preprocess the image m

b. Creating-dictionary (m)

c. Feature-extraction (m)

d. K-means (data, n)

e. SVM-training (samples, label)

f. Classify-image (n)

g. Classification-result (n)

End

Creating-dictionary (m) {

h. Feature-extraction (m)

i. SIFT-key (m)

j. KAZE (m)

k. Generate feature vector

l. Apply PCA }

K-means (data, n) {

m. Set initial centers of clusters, c_1, c_2, ….., c_k to the arbitrarily selected $k$vectors

n. Classify each vector $x_1 = [x_{11}, x_{12}, \ldots, x_{1d}]$ into the closest center c_i

o. Recalculate the cluster center $c_i = [c_{i1}, c_{i2}, \ldots.., c_{id}]$ until centroid no longer moves

p. Assign the label.

q. }

SVM-training (samples, labels)

build SVM-kernel classifier

classify-image (n)

input image for classification

match (n)

(Continued)

Algorithm: (continued)

//extract local descriptor for testing image

i. SIFT-key (n)

ii. KAZE-key (n)

Assign the descriptor to visual modeling

Compare images using SIFT-key and KAZE-key

Classification-result (n)

//classify image using SVM

SVMClass (Test Feature),

Display classification result

6 Experimental Results and Discussion

6.1 Experimental Setup

The required simulation is performed using the MATLAB software, the 64-bit operating system of Windows 7 that includes the computational resources of RAM 4 GB, CPU Intel Core i5 with an operating frequency of 3.1 GHz.

6.2 Dataset

To evaluate the performance of the proposed method the versatile standard datasets such as Caltech 256 and Corel 10 K datasets are used. Tab. 1 shows the various dataset and their details. Caltech 256 dataset is the improved version of Caltech 101 with a lot of improvement. The categories of Caltech dataset comprises of "Buddha", "Cake", "Bonsai", "Butterfly"," Wrist watch" and "Leopard". The Corel image dataset contains rich content images with a generous range. The categories of Corel dataset comprises of "Butterflies", "Food", "Dinosaurs", "Roses", "Postcards", "Air Balloons", "Monkeys", "Flowers", "Pills", "Buses", "Furniture", and "Caves".

Table 1: Various datasets used in this approach

Data sets	Number of categories	Number of images	Size of the image	URL
Caltech 256	256	30,607	300 × 200 pixels	http://www.vision.caltech.edu/Image_Datasets/Caltech256/
Corel 10 K	100	10,000	192 × 128 pixels	http://wang.ist.psu.edu/docs/related/

The corresponding sample images are shown in Figs 4 and 5. In Corel 10 K dataset for training 70% images are selected and for testing 30% images are selected. In Caltech 256 dataset for training 60% images are selected and for testing 40% images are selected. This ratio was chosen to have more training data that result in the well-formed codeword of the vocabulary, which results in a stable solution. The K fold cross-validation method is selected to calculate the accuracy and performance of the visual descriptor. It divides the training dataset into smaller equal subsets i.e., K folds, and run the classification for K number of times, each time picking a different subset. Each experiment is performed by choosing the value of K as 10.

The image retrieval performance is done on the test image dataset. The closeness of the classifier score values is calculated to retrieve the images. The image class is determined by the classifier output label. The Euclidean distance is used between the scores of the images stored in an image database and the score of a given query image to validate the output of retrieved images. The selected feature percentages are varied to reduce the computational cost. It is studied that an increase in the size of the dictionary increases the performance of the image retrieval. To achieve the best results dictionary size is varied (i.e., 100, 200, 300, 400, 600, 800, 1000, and 1200).

Figure 4: Sample images from Caltech 256 dataset

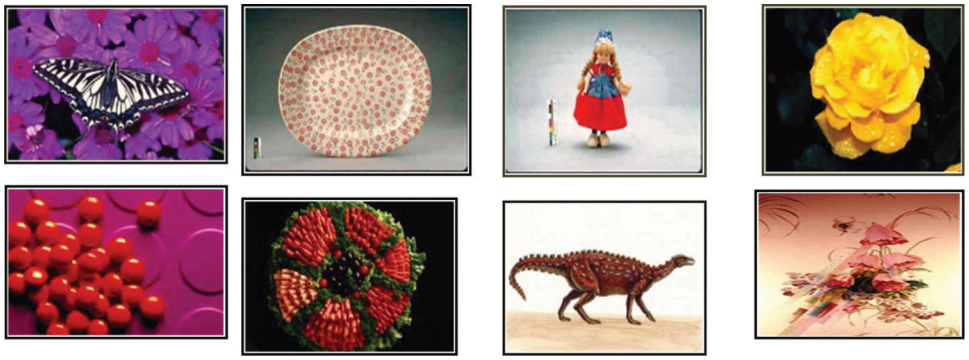

Figure 5: Sample images from Corel 10 K dataset

6.3 Evaluation Parameter and Metrics

In this work, we consider dictionary size and feature percentages per image as two important parameters which influence the performance of the CBIR. In order to evaluate the best performance of the proposed technique, dictionaries of different sizes are used with different features percentages (i.e., 10%, 25%, 50%, 75%, and 100%) per image. The precision determines the number of correctly retrieved images over the total number of retrieved images from the test image database. It measures the specificity of an image retrieval system, represented as:

$$\text{Precision} = \frac{\text{Total no. of. correctly retrieved images}}{\text{Total no. of. retrieved images}} \tag{7}$$

The ratio of correctly retrieved images over the total number of relevant images of that semantic class in the image database is known as recall and it measures the sensitivity of the image retrieval system, represented as:

$$\text{Recall} = \frac{\text{Total no. of. correctly retrieved images}}{\text{Total no. of. relevant images in the database}} \tag{8}$$

The performance is measured by mean average precision and accuracy. The performance analysis measured in terms of the mean average precision (mAP) of the proposed method is presented in Tabs. 2–4 based on visual words fusion on different sizes of the dictionary using various feature percentages.

Table 2: mAP performance based on visual words fusion on different dictionary sizes and feature percentages on Corel-10 K database

Selected feature	mAP analysis based on different dictionary sizes							
	100	200	300	400	600	800	1000	1200
25%	71.63	73.45	74.58	76.62	77.52	79.68	83.99	84.23
50%	76.21	77.68	76.91	78.91	79.96	82.11	84.72	84.98
75%	78.32	79.32	81.89	82.87	83.87	86.22	89.94	90.12
mAP	75.38	76.81	77.79	79.46	80.45	82.63	86.21	90.55

Table 3: mAP performance based on visual words fusion on different dictionary sizes and feature percentages on Caltech-256 database

Selected feature	mAP analysis based on different dictionary sizes							
	100	200	300	400	600	800	1000	1200
25%	77.13	78.12	79.11	81.98	82.61	83.68	84.99	85.62
50%	79.21	80.68	81.93	83.94	84.96	85.23	86.72	86.74
75%	83.28	84.78	85.91	86.58	86.78	88.39	89.52	89.81
MAP	79.87	81.19	82.31	84.16	84.78	85.76	87.07	87.39

Table 4: Comparison of mAP performance with other methods based on different dictionary sizes for the feature percentage 75 for the Caltech–256 datasets

mAP value	Dictionary size							
	100	200	300	400	600	800	1000	1200
SIFT & FREAK based on BoVW	71.13	73.99	74.12	74.55	74.87	74.25	73.28	76.02
SIFT & LIOP based on BoVW	74.1	77.2	77.92	78.12	79.24	82.9	74.64	76.04
*Proposed method (SIFT & KAZE based on BoVW)	79.87	81.19	82.31	84.16	84.78	85.76	87.07	87.39

The Tabs. 2 and 3 shows the performance of selected feature size of Corel 10 K and Caltech-256 on different dictionary sizes. The experimental results are analyzed and shown in Figs. 6–8. According to the experimental results shown in Tab. 3, the proposed method confirms finest performance than the other existing research method. With the intention of observing the best performance of the proposed method based on fusion of SIFT and KAZE descriptors, different sizes of the dictionary (i.e., 100, 200, 300, 400, 600, 800, 1000, and 1200) using different features percentages (i.e., 25%, 50%, 75%, and 100%) per image are used. The motive for selecting different features percentages per image is to reduce the computational cost. As per the codebook algorithm, Dictionary size or cluster size influences the retrieval performance. We observed that in this method as the dictionary size is increased, the ability of dictionary to identify different images is strengthened. However, the proposed work shows better retrieval performance for Caltech-256 and Corel10k in case of complex background, clutter and overlapping images and they also provide promising results for the other categories too. Further to validate the robustness of the proposed technique, the mAP performance analysis is compared with the existing SIFT-FREAK technique and SIFT-LIOP technique which is shown in Fig. 8. It clearly indicates that the proposed technique based on fusion of SIFT-KAZE yields better performance as compared to the existing CBIR technique. Another interesting behavior that can see in our work is the retrieval of spatial details of large-scale salient images without any segmentation process, which enhances the scalable performances. By employing PCA and bag of visual words in the proposed method, the over fitting problem caused by the large dictionary size is reduced. The best mAP performance is attained using the proposed method based on features fusion of SIFT and KAZE features which is 87.07% with a dictionary size of 1000 and using 75% features per image.

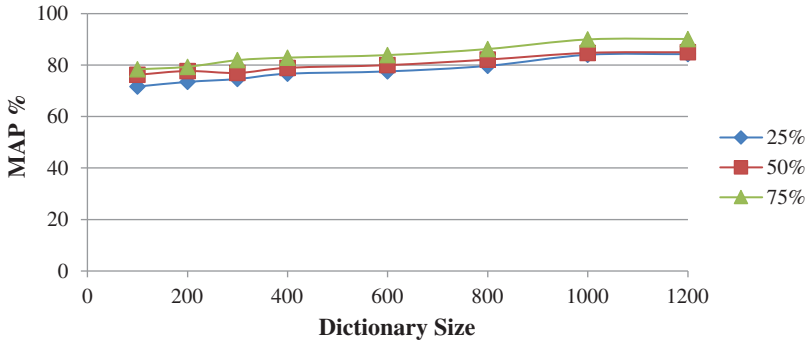

Figure 6: mAP evaluation on different sizes of the dictionary on the Caltech-256 database

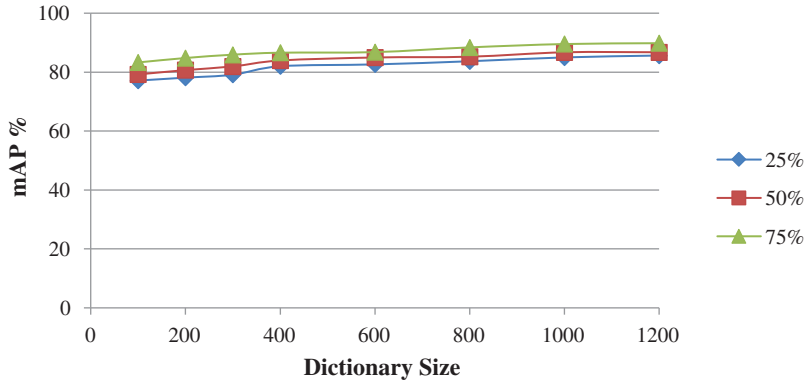

Figure 7: mAP evaluation on different sizes of the dictionary on the Corel 10 K database

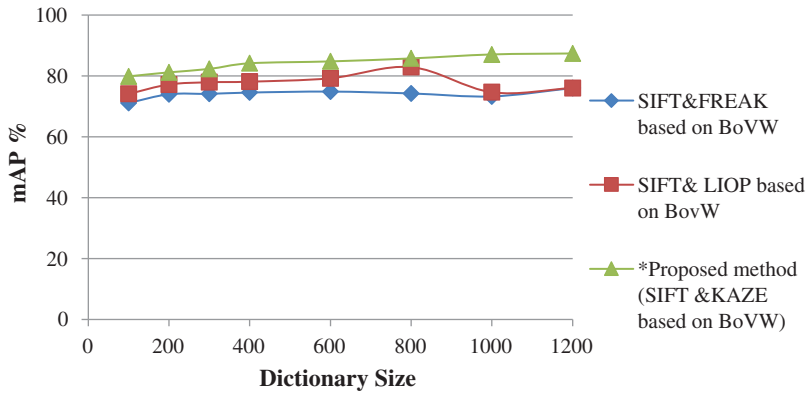

Figure 8: Comparison of mAP evaluation of existing research methods with our proposed method

7 Conclusion

In CBIR, the image descriptor plays a vital role in assessing the similarities among images. The proposed method based on feature fusion of SIFT and KAZE enhances the scalability of CBIR. PCA employed in this method eliminates the problem of over fitting by reducing the feature dimension. RF employed in this method reduces the semantic gap between high-level and low-level features, while preserving the original topology of the high-dimensional space. It is observed that the fusion methodology of SIFT and KAZE feature descriptors overcomes the multi-scale issues. Performance comparison shows that the proposed method produces better means average precision. In the future, we plan to deploy pre-trained convolutional neural networks.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-based image retrieval at the end of the early years," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] C. Su, H. Chiu and T. Hsieh, "An efficient image retrieval based on HSV color space," in *Proc. of Int. Conf. on Electrical and Control Engineering*, Yichang, China, pp. 5746–5749, 2011.

[3] J. Ma, "Content-based image retrieval with HSV color space and texture features," in *Proc. of Int. Conf. on Web Information Systems and Mining*, Shanghai, pp. 61–63, 2009.

[4] G. J. Scott, M. N. Klaric, C. H. Davis and C. Shyu, "Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 5, pp. 1603–1616, 2011.

[5] S. Belongie, J. Malik and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[6] J. R. Smith and S. Chang, "Automated binary texture feature sets for image retrieval," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing Conf. Proc.*, pp. 2239–2242, vol. 4, Atlanta, USA, 1996.

[7] H. A. Moghaddam and M. Saadatmand-Tarzjan, "Gabor wavelet correlogram algorithm for image indexing and retrieval," in *Proc. Int. Conf. on Pattern Recognition*, Hong Kong, pp. 925–928, 2006.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] H. Bay, A. Ess, T. Tuytelaars and L. V. Gool, "Speeded-up robust features SURF," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[10] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. on Computer Vision*, Barcelona, pp. 2564–2571, 2011.

[11] R. Datta, D. Joshi, J. Li and Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–6, 2008.

[12] A. Halawani, A. Teynor, L. Setia, G. Brunner, and H. Burkhardt, "Fundamentals and applications of image retrieval: An overview," *Datenbank-Spektrum*, vol. 18, pp. 14–23, 2006.

[13] S. A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. Ninth IEEE Int. Conf. on Computer Vision*, France, vol. 2, pp. 1470–1477, 2003.

[14] J. J. Rocchio, "*Document retrieval system: Optimization and evaluation*," Ph.D. dissertation, Harvard Computational Lab, Harvard University, Cambridge, MA, 1996.

[15] A. Vadivel, S. Shamik and A. K. Majumdar, "An integrated color and intensity co-occurrence matrix," *Pattern Recognition Letters*, vol. 28, no. 8, pp. 974–983, 2007.

[16] K. van de Sande, T. Gevers and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[17] A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognition*, vol. 29, no. 8, pp. 1233–1244, 1996.

[18] C. Jin and S. W. Ke, "Content-based image retrieval based on shape similarity calculation," *3D Research*, vol. 8, no. 23, 2017.

[19] M. Aly, P. Welinder, M. Munich and P. Perona, "Automatic discovery of image families: Global vs. local features," *16th IEEE International Conference on Image Processing (ICIP)*, pp. 777–780, 2009.

[20] F. Yang, Y. Xu and H. Shen, "Many local pattern texture features: Which is better for image-based multilabel human protein subcellular localization classification," *Scientific World Journal*, pp. 1–14, 2014.

[21] N. Bayramoglu and A. A. Alatan, "Shape index sift: Range image recognition using local features," *20th International Conference on Pattern Recognition*, Istanbul, Turkey, pp. 352–355, 2010.

[22] G. Csurka, C. R. Dance, L. Fan, J. Willamowski and C. Bray, "Visual categorization with bags of key points," *Workshop on Statistical Learning in Computer Vision*, vol. 1, Meylan, France, pp. 1–22, 2004.

[23] Y. Liu, D. Zhang, G. Lu and W. Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[24] H. Jegou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.

[25] P. S. Hiremath and J. Pujari, "Content based image retrieval using color, texture and shape features," in *Proc. of Int. Conf. on Advanced Computing and Communications*, Guwahati, Assam, pp. 780–784, 2007.

[26] R. Ashraf, M. Ahmed, S. Jabbar, S. Khalid, A. Ahmad, "Content based image retrieval by using color descriptor and discrete wavelet transform," *Journal of Medical Systems*, vol. 42, no. 3, 2018.

[27] K. T. Ahmed, S. Ummesafi and A. Iqbal, "Content based image retrieval using image features information fusion," *Inf. Fusion*, vol. 51, pp. 76–99, 2019.

[28] H. Bu, N. Kim, C. Moon and J. Kim, "Content-based image retrieval using combined color and texture features extracted by multi-resolution multi-direction filtering," *Journal of Information Processing System*, vol. 13, pp. 464–475, 2017.

[29] E. G. Karakasis, A. Amanatiadis, A. Gasteratos and S. A. Chatzichristofis, "Image moment invariants as local features for content-based image retrieval using the Bag-of-visual-words model," *Pattern Recognition Letters*, vol. 55, pp. 22–27, 2015.

[30] Z. Mehmood, S. M. Anwar, N. Ali, H. A. Habib and M. Rashid, "A novel image retrieval based on a combination of local and global histograms of visual words," *Mathematical Problems in Engineering*, vol. 2016, pp. 1–12, 2016.

[31] Z. Mehmood, T. Mahmood and M. A. Javid, "Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine," *Applied Intelligence*, vol. 48, no. 1, pp. 166–181, 2018.

[32] I. Dimitrovski, D. Kocev, S. Loskovska and S. Dzeroski, "Improving bag-of-visual-words image retrieval with predictive clustering trees," *Information Sciences*, vol. 329, pp. 851–865, 2016.

[33] G. Swathi Rao, "Effects of image retrieval from image database using linear kernel and hellinger kernel mapping of svm," *International Journal of Scientific & Engineering Research*, vol. 4, no. 5, pp. 1184–1190, 2013.

[34] W. X. Liu, J. Hou and H. R. Karimi, "Research on vocabulary sizes and codebook universality," *Abstract and Applied Analysis*, vol. 2014, pp. 1–7, 2014.

[35] S. Maji, A. C. Berg and J. Malik, "Efficient classification for additive kernel svms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 66–77, 2013.

[36] M. Alkhawlani, M. Elmogy and H. Elbakry, "Content-based image retrieval using local features descriptors and bag-of-visual words," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 9, pp. 212–219, 2015.

[37] U. Sharif, Z. Mehmood, T. Mahmood, M. A. Javid, A. Rehman, "Scene analysis and search using local features and support vector machine for effective content-based image retrieval," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 901–925, 2019.

[38] D. Zhang, M. M. Islam and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognition*, vol. 45, no. 1, pp. 346–362, 2012.

[39] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, pp. 100–108, 1979.

[40] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[41] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 480–492, 2012.

[42] J. Li and N. M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing*, vol. 71, no. 10, pp. 1771–1787, 2008.

[43] M. Abdullah, S. A. Khan, M. Alenez, K. Almustafa and W. Iqbal, "Application centric virtual machine placements to minimize bandwidth utilization in datacenters," *Intelligent Automation and Soft Computing*, vol. 26, no. 1, pp. 13–25, 2020.

[44] P. F. Alcantarilla, A. Bartoli and A. J. Davison, "KAZE features," in *Proc. of European Conf. on Computer Vision 2012*, pp. 214–227, 2012.

[45] S. A. K. Tareen and Z. Saleem, "A comparative analysis of sift, surf, kaze, akaze, orb and brisk," in *Proc. of Int. Conf. on Computing, Mathematics and Engineering Technologies*, Sukkur, Pakistan, pp. 1–10, 2018.

[46] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, pp. 327–352, 1977.

[47] Y. Rui, T. S. Huang, M. Ortega and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644–655, 1998.

[48] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Systems*, vol. 8, no. 6, pp. 536–544, 2003.

[49] R. Fernandez-Beltran and F. Pla, "Latent topics-based relevance feedback for video retrieval," *Pattern Recognition*, vol. 51, pp. 72–84, 2016.

[50] S. Rota Bulo, M. Rabbi and M. Pelillo, "Content-based image retrieval with relevance feedback using random walks," *Pattern Recognition*, vol. 44, no. 9, pp. 2109–2122, 2011.