Khan et al. LGURJCSIT 2023

ISSN: 2521-0122 (Online)
ISSN: 2519-7991 (Print)

LGU Research Journal of Computer Science & IT

doi: 10.54692/lgurjcsit.2023.074532
Vol (7): Issue (4), October – December2023

Efficient and Accurate Image Classification Via Spatial Pyramid Matching and SURF Sparse Coding

Saif Ur Rehman Khan [1*], Asif Raza [2], Muhammad Waqas[3], Muhammad Abdur Raphay Zia[4]
[1]School of Computer Science and Engineering, Central South University, Changsha, Hunan, China.
[2]Department of Computer Science, Bahauddin Zakariya University Multan, Punjab, Pakistan.
[3]FAST Schools of Computing, Department of Artificial intelligence and Data Science, National University of Computer and Emerging Science (FAST-NUCES), Karachi, Pakistan.
[4]Department of Computer Science, University of Central Punjab, Lahore, Pakistan.

Email: saifurrehman.khan@csu.edu.cn

ABSTRACT:

Sparsely coded signal space representations do well in feature quantization. Instead of using standard vector quantization, the suggested method uses selective sparse coding to assemble the most important features of the appearance descriptors of nearby image patches. Inadequate coding also enables adjacent max pooling on some spatial scales, which, unlike the setup of average pooling in a histogram, links interpretation with scale invariance. The acquired visual illustration is the key contribution of this research; it performs outperform with linear-SVM, improves the model training's, which in turn speeds up testing with improves accuracy. The efficacy of the method we have employed has been substantiated through a series of experiments conducted on diverse datasets. Since top-performing image classification systems heavily rely on nonlinear SPM in mean of vector quantization, the trustworthy recommended linear SPM greatly increases the use of larger sets of training data. The method given herein deduces that the sparse coding of SURF feature's function hampered a more comprehensive local appearance descriptor for general-purpose image processing. Experiments and comparisons are conducted on standard datasets such as Caltech-101, FTVL, and Corel-1000, using state-of-the-art techniques and descriptors. When compared over several other image categories and descriptors, the method provided here comes out on top.

KEYWORDS: Machine Learning, Classification, SURF Sparse Coding, Spatial Pyramid, Image Dataset.

1. INTRODUCTION

Massive collections of digital photographs are being generated in many sectors, including government, business, healthcare, and academia. By scanning in existing collections of analog artworks like schematics, images, paintings, prints, and sketches, these new digital collections have been created. Image and object recognition, texture categorization, scene understanding, and symmetry detection are all areas of computer vision that have benefited from research contributions. Classifying a digital image into categories such as "water," "forest," "flower," "crops," "soil," "vehicle," etc. While it may be a straightforward task for humans, mastering this process has proven to be a formidable challenge for computers. An image is the only kind of picture that can convey meaning. Photographs of people, publics, animals, the outdoors, and interesting sights; microphotographs of electrical components; and medical imaging results. In a basic sorting system, a camera positioned high above the region of interest takes images that are then processed. A random blur is unfair, even if the image is unrecognizable. Classification systems use databases containing predetermined

patterns to assign items to appropriate categories based on how closely those patterns match the characteristics of the recognized object. An image of the object is provided as a query image in order to determine whether or not it belongs to one or more categories of interest. An object's perspective, as much as the thing itself, can be gleaned from its classification. Classification of digital images is performed to manage each image in accordance with its category [1]. In order to organize a huge collection of photographs. In the study of computer vision, an image's context is crucial to the classification process. A digital image can be classified into multiple categories based on the objects it contains. Therefore, images are mined for categorization purposes in order to extract relevant data for classifier use. Several methods are used to extract features from digital images, and the classifier then sorts images into categories based on those attributes.

Classifiers, that rely on attributes can differentiate between objects based on their motion or shape [2]. Shape-based categorization strategies often make use of attributes such as contour, silhouette, skeleton, points, and primitive geometric shapes, while motion-based strategies have traditionally relied on attributes such as object motion. Training sample-based algorithms makes use of both supervised and unsupervised classification methods. Supervised classification methods use samples from known information categories to train on, allowing for the presentation of relevant orientation statistics prior to classifying unknown pixels. The classification algorithm is then primed using the signals generated by the training sample set to classify the spectral information into a map. In contrast, unsupervised categorization relies on the statistical information already present in the image to investigate and separate a large set of unseen pixels into smaller groups, with no need for extensive background knowledge.

Parametric and non-parametric classifiers are examples of parameterized techniques to data analysis. Parametric classifiers, which are created from training samples and assume a Gaussian distribution, use parameters like the covariance matrix and the mean vector [8]. However, in non-parametric classifiers, no statistical parameters are used in the computation of class separation. Pixel information based classifiers. Object oriented classifiers consist of two phases: image segmentation (the act of grouping image pixels into objects) and classification (the act of doing classification based on the grouped objects). Signature in per-pixel classifier is created by combining feature bands of all training probed pixels. Assuming that every pixel is associated with every end member. Soft and hard classifiers are two examples of classification strategies that use production amounts for each geographic component as their basis. In hard classification, each pixel must demonstrate membership in exactly one class, whereas in soft classification, membership in many and/or partial classes can be demonstrated by any given pixel. Spatial information can be used in a number of different classification approaches [3-5] Classifiers. Non-parametric and parametric classifiers produce initial classification images, and then contextual classifiers are applied on classified images, Contextual image classification surpasses the utilization of merely spatially adjacent pixel information and spectral data, which are the primary inputs for spectral classifiers. Kernel-based methods are those that solve linear problems in a higher-dimensional space by mapping the kernel feature space to the original feature space, and this allows for a more straightforward geometric interpretation of learning algorithms. A feature of interest can be represented by a feature vector. Changing the perspective or interest point detection algorithm can have a significant effect on the detected features. Extracting a ROI from a digital image and then using its features in other contexts is a common example of low-level feature vector obtaining. A texture element, often known as a Texel, is a feature that is present in all areas of the same texture [6]. Texture features contain useful spatial information for discrimination. Images of walls, or patterns on cloth or other surfaces stand out, and the ability to detect texture through surveillance is dependent on factors like as viewing angle, lighting, distance, and other environmental impacts. Texture features can be thought of as a collection of local statistical properties of the pixel gray level strength [6], and are used in the interpretation and analysis of digital images. Fine textures have a large number of edges or spatial frequencies per unit area, while coarse textures contain a small number of edges or low spatial frequencies [7]. Texture analysis comprises four distinct procedures: texture extraction, classification, segmentation, and shape-from-texture [8]. There are many methods, such as Markov random fields, Gabor filters, and

co-occurrence matrices, for extracting texture features [9]. Graph based techniques, which effectively depict the relationship between pixels, are also useful for picture segmentation applications [10,15]. Color features accurately describe the visual content of digital images, and even the most basic color extraction can improve image classification [16]. Edges can be indicated by form features, which are less affected by variations in lighting, but which do vary with object orientation and size. Classification of the object is possible on the basis of shape [17-20]. Divided image into multiple regions using various techniques, and within each zone are many items. While shape characteristics are good at identifying areas in an image, they need color features to create a complete description of the image. All areas are taken from photographs, and each one is made up of a collection of pixels that are then reflected in a new image.

Numerous classifications can be fetched from the large image and class collection. It is a critical effort for organizing digital photographs to classify thousands of images into distinct categories. Common features used for classification are those present in the images themselves. item identification segmentation based on obvious borders of item or color variation between object is commonly used to separate foreground objects from background objects in modern picture classification algorithms. In order to approximate similarity between feature vectors, spatial pyramid matching is used. For each feature space, spatial pyramid matching superimposes a series of grids with progressively finer resolution, computing a weighted total of the number of matches at each level of resolution; two points are considered equivalent if they occur in the same level grid. Hence, the combination of Spatial Pyramid Matching alongside SURF sparse coding yields superior results in feature extraction and image classification tasks.

Images can be represented by a set of local image descriptors, which can then be used for image search and categorization based on criteria such as shape, color, texture, and areas. Despite its importance for learning-based classifiers, the quantization of local descriptors is damaging and wasteful in non-parametric classifiers because there is no training phase to compensate for the lost data [20]. It's been intriguing to think about selecting from a set of feature mining techniques to obtain a descriptor that's unique to a given image class or image in order to acquire a more exact description of the image's content. Consider the portrayal of images while taking into account the organization of their associated points, especially if we conceptualize a descriptor as a point within a high-dimensional component space. We hope that the descriptors, and the distribution of their associated focuses in the feature space, will reflect our expectation that images that have been placed in the same class will be more similar to one another than images that have been placed in different classes. Each possible descriptor configuration strives to provide a better, more robust means of encoding these similarities within the descriptor. Introductions and examples of use for HOG [21]., DoG [22], and LBP [23], are provided below. For the sake of classification or other machine learning tasks, these descriptors characterize the visual content of digital photographs. Each cell in the pyramidal grid is assigned a bag of words in the spatial pyramid representation method [24]. Assembling Bag of Word data from neighboring cells to create a spatial pyramid representation, which deconstructs images at multiple levels in a recursive fashion. To determine how many matches there are, the histogram crossing point kernel is calculated between the two connected regions. Coordination at a finer granularity has less of an effect than coordination at a coarser granularity when a penalty weight is applied. When it comes to the distribution of nearby features, the more matches there are, the closer they are. Therefore, the spatial pyramid representation is a good option if scenes and objects frequently occur that require significant image adjustments. Spatial graphs, where nodes represent blocks of spatial data in SPM and edges represent relationships between these blocks, are used to address the fact that SPM disregards the spatial information included in their relationships [25]. The fundamental objective of image classification is to categorize an image into one or more predefined semantic categories.

• By employing a combination of spatial pyramid matching, SURF sparse coding, spatial max pooling, and SVM techniques, you can develop an innovative approach to image classification.

• The method improves classification efficiency while cutting down on training time by decreasing the feature vector size.

2. METHODS AND MATERIALS

First, a set of interest points is selected from the

digital image in a random square pattern. The feature vector is then used to depict where each interest point is, and blob detection is performed on the basis of interest point detection. The retrieved feature vector is then classified using a SVM classifier, after which scale space division and interpolation are performed for efficient feature extraction using sparse coding. Figure 1 has presented the proposed methodology architecture.

2.1. Query Image

The classification system receives a query image as input and then classifies it after applying various transformations to it, such as gray-level conversion, noise removal, interesting point detection, blob detection, scaling the image, dividing the scale space, interpolating the scale space, applying sparse coding, and spatial max pooling.

2.2. Conversion (RGB to Gray)

RGB image pixel values for red, green, and blue are 190, 183, and 175. These R, G, and B channels each start with an initial pixel value of (190, 183, 175) at (1, 1). Each pixel can have a strength between 0 and 255, where 0 is black and 255 is white. This optimization boosts color depth and allows for more accurate grayscale computation. Because of their computational simplicity, binary images are often used for image enhancement and edge detection, prompting a widespread shift toward binary image formats. Images in color can be transformed into monochrome ones. The equation can be used to complete the transformation [26].

$$Kx=0.333\ Kr + 0.5\ Kg + 0.1666\ Kb \qquad (1)$$

Where Kr, Kg, and Kb are the respective R, G, and B factor strengths and Ky is the strength of the RGB image's equivalent gray level images. The grayscale contrast during color-to-grayscale conversion should obviously mimic the color contrasts. Grayscale contrasts with a negative or positive divergence of gray level varies should naturally correspond to color contrasts with a corresponding divergence of brightness varies. The grayscale image's tunable series of gray levels should naturally coincide with the color image's tunable range of brightness values. A constant function is used in the transformation from color to grayscale. In a grayscale image, if two adjacent pixels have the same color, they will also have the same gray level. If a pixel in the original color image is gray, the corresponding pixel in the grayscale version will also be gray, allowing you to establish a connection between brightness and tonality. In the same way that a series of pixels with increasing luminance in a color image will have the same saturation and hue, a series of grayscale pixels with increasing gray levels will have the same saturation and hue, minimizing image artifacts. In the first phase of the proposed method, a colored query image is converted to a grayscale image. In order to effectively remove noise from a digital image, many methods are used to the gray level image. The primary reason for converting a color image to a grayscale image is to reduce the amount of data included in each individual pixel; this is necessary since color images are more difficult to analyze and comprehend, making grayscale images the only option for many tasks, including classification. If the levels of gray are evenly spaced, the human eye is far better at distinguishing between them than the variance among continuous gray levels.

2.3. Filtering (Square-Shaped)

In order to quickly calculate box-type filters, grayscale images are integrally represented. Integral images can have noise removed by using a square-shaped filter. Digital images with noise, which is a subjective difference in color details or brightness that obscures expected information, can be difficult to interpret. As a result, noise can be filtered out using Gaussian smoothing. The squared potential zones of interest in the query image are detected by the Gaussian smoothing. In addition, a 9-16 mask has been developed, which calls for a minimum of a 9-pixel circle within a 16-pixel square to be brighter than the center pixel. Additionally, Hessian matrix estimate is utilized for digital image landmark detection. Integral images can be formed more efficiently using a box-type filter by employing square-shaped filtering at both the octave and intra-octave levels. The total of all pixels at a given point p= (x,y)l within a rectangular segment of the input image I is denoted by the integral image Figure 1. Calculated integral images [26] use the three additions from eq. (1) to determine the total intensity.

$$II_{\Sigma}(\boldsymbol{p}) = \sum_{m=0}^{m\le x} \sum_{n=0}^{m\le y} II(m,n) \qquad (2)$$

2.4. Hessian Matrix Based Interest Point Detection

Hessian affine region detector is a good performer in terms of accuracy and computation time, and it is based on corner detection, such as areas with low self-similarity and change in light intensity, and auto correlation, which is used to specify the key points in image. We push the estimation substantially further using box filters because Gaussian filters are imperfect in a y scenario and because we are confident in Lowe's success with Laplacian of Gaussian approximations. Integral images of any size can be used for a fast evaluation of these estimated second-order Gaussian derivatives.

2.5. Determinant (Blob Detector)

Scale derivatives are calculated using box filters in the current iteration, and the derivatives are then smoothed using a Gaussian kernel. The weights are applied to square sections where calculation time is constant regardless of filter size, hence reducing the computational cost. The 99 strongbox filters with K=1.2 Gaussian estimates reveal the smallest possible scale for generating blob response maps. Blobs are simple, low-level objects that can take on any shape and dimension. In order to obtain a pyramid of interest points at varying scales, several blob detection methods rely on a scale-space representation of the image. To improve performance and gain contextual knowledge about regions, blob detectors are applied to squared regions. Points of illustration are typically dispersed alongside the pixel grid of an image, and the value of each pixel is used to determine the point's illustration value. As a result, the sample weight is set so that foreground pixels have a heavy weight and background pixels carry a light one. In order to create a multi-scale image description, we interlace the image with expanding Gaussian filters of spatial change. This creates a scale space description with two spatial directions and a third direction displaying scale. By weaving the image with Gaussian channels of increasing spatial change, we can create a scale-space representation of the image, in which the two spatial dimensions are accompanied by a third dimension that speaks to scale or determination.

2.6. Scale Space Division

When upscaling an image, the entire scale space is separated so that the scale and space of image pyramids can be accumulated. Changing the values of these factors, which stand in for the objects they depict, yields a variety of visual results. Scale space division is a method used in image analysis to simplify images so that more accurate findings can be obtained while maintaining or improving computational speed. To obtain a higher-level pyramid, we subsample the images and smooth them with a Gaussian distribution. The speed at which larger masks can be applied to the original image's box filters is unaffected. For a factor-of-two scale shift, the resulting filters have dimensions of 9x9, 15x15, 21x21, and 27x27. Scaling up requires a corresponding increase in the size of consecutive filters. In order to display the results of a series of filters, octaves are created by the split of scale space. The number of semitones in an octave remains constant, while the octave itself incorporates a scaling factor of 2. Because integral images are discrete, the difference between any two consecutive scales depends on the length by a factor of 10 in the direction of the second order partial derivative (i or j) of the positive or negative lobes. Only the first and last Hessian response maps are used for this comparison. The maxima of the determinants of the Hessian matrix are interpolated to the image's scale and space.

2.7. Interpolation

Scale-space interpolation is used to improve an image's semantic content by combining information from neighboring pixels. Interpolation in the scale space is used to make calculations between pyramid samples. After interpolation, the smallest feasible scale K = 1.6 = 1.2 12/9, which is the same as the size of a 12 x 12 filter, and the largest is K = 3.2 = 1.2 24/9. Each additional octave requires twice the size of the filter, so this principle applies to them as well. The sample interval is also doubled to cut down on the accuracy loss and calculation time for interest point extraction. The sizes of the filters used in the second octave are 15, 27, 39, and 51, and the sizes used in the third octave are 27, 51, 75, and 99. If the input image size is not quite equal to the matching filter sizes, then a scale space analysis will be required for the fourth octave. The fourth octave uses filters with numerical values of 51, 99, 147, and 195. Other octaves can be determined in the same way. The comparatively coarse sampling at these scales is a result of the changes at the macro level. To begin the first

octave, a filter of size 15 is used, and to begin the second octave, the filter size is increased by an additional 12 pixels. This will adjust the ratio between the first two filters to 1.4 * (21/15). Using quadratic interpolation, the smallest detectable scale for the exact version is s=((1.2 18/9))/2=1.2.

2.8. Sparse Coding

Typically, SURF descriptors are both sparse in the transverse direction and spatially explicit in the same image. It is possible to represent input vectors as a linear combination of weights by employing sparse coding. The primary goal of sparse modeling is to build a vocabulary D where $X \simeq DV$ with $\|v_i\|_0$ for most x_i in the data, where v_i is suitably minimal. For simplicity, we'll refer to D as a vocabulary of U atoms as $d_u \epsilon R^m$., and X as a collection of N column data vectors $x_i \epsilon R^m$, $D \epsilon R^{m \times U}$. Restoration constants $v_i \epsilon R^U$ will be aligned along the columns of a matrix for each data vector x_i.

$$V= [v_1,...,v_N] \epsilon R^{U \times N} \quad (3)$$

Sparse coding [27] refers to the process of calculating V for a given D. Equation (4) is a representation of the common l_0 or l_1 consequence modeling problem [28].

$$(V^*, D^*)=\arg \min_{V,D} \|X- DV\|_F^2+ \lambda \|V\|_p \quad (4)$$

Where and p=0,1 and $\|.\|_F$ is the Frobenius norm. For each column in V, the rate function includes a quadratic *fitting term* and a l_0 or l_1 regularization term, with the balance between the two determined by the penalty parameter. It is possible to estimate l_0 using the l_1 benchmark. It is common practice to use l_0 penalty for reconstruction and l_1 penalty for categorization. The sparse representation of picture features that is returned by sparse coding can boost discrimination and classification accuracy. To improve performance and outcomes, sparse coding is applied to feature vectors, yielding sparse feature vectors.

2.9. Spatial Pooling

Let's pretend that every digital image is represented by a set of descriptor vectors v that spreads in a way unique to that image, as shown by the density function density (v) for a background measure that is independent of the image itself, dμ (v). v's location data was integrated via spatial pyramid matching. Spatial pooling is used to calculate local features that are then used to represent images. Since max-pooling is used to obtain extreme features in sparse coding, the suggested method improves feature discrimination and classification accuracy. To get better performance and more impressive results in image classification, spatial pooling segments images into many blocks at multiple levels and then calculates feature vectors for each segment. The entire collection of digital images is then organized into multiple degrees of subdivisions, and pooling is applied to each of those subdivisions based on the blocks from which each feature descriptor was originally extracted. The results of the local pools are then combined to obtain the spatial pyramid representation of the digital image.

$$s_{i,j} = \max_{m \epsilon ni} x_{m,j} \quad \text{For } j=1,\dots,Z \quad (5)$$

Maximum spatial pooling of picture interest regions is represented by eq. (4). Better results can be achieved in classification to the spatial arrangement of images, which contains more reliable and efficient information.

2.10. Classification

The classifier is responsible for classifying images, and it can use either supervised or unsupervised methods to do so. SVMs, being supervised learning models with associated learning methodologies that assess image characteristics utilized for classification, were the ones we had chosen for this purpose. In order to create a non-probabilistic binary linear classifier from the training data provided, the SVM training approach creates a model that assigns new samples to one of the two classes. The samples in an SVM framework are points in space that are connected as the samples from each class are clearly differentiated from one another. Then, new samples are added to the pool and expected to fit into a preexisting classification scheme based on their side of the difference. Classifying images is accomplished with the assistance of classifiers using the image representations acquired from the spatial pooling step. This is followed by the classification of similar images using a support vector machine classifier. In order to do classification efficiently and with a higher degree of accuracy, max spatial pooling extracts more robust information from images. To train, we employ a linear support vector machine with a

hinge cost. Classification performance at the picture level is inextricably bound to the actual value of performance on local descriptors. Let's assume, for the sake of argument, that only global pooling is used, in which case a patch-level pattern correspondence is performed in all digital image space, and the results are then accumulated to produce the groove showing how reliably available a certain class of object is likely to be. For the most part, biologically-inspired vision frameworks use feed-forward pattern matching for visual object detection, and this insight works in harmony with that methodology. This group emphasizes the need of learning a good encoding scheme for local feature descriptors, which ultimately determines whether or not the elusive classification algorithm can be learned successfully. The group also thinks that supervised trainings of possibly will lead to further improvements.

3. RESULT AND IMPLEMENTATION

Converting color space to grayscale is the first step toward fast computing. By processing the grayscale image, local interest points based on intensity are discovered, and feature vectors are generated. Better interpretation of features is achieved through the extraction of global features for the interest sites using an improved sliding window. The retrieved characteristics and texture features are then merged for effective picture classification. Coefficients of restructured observations are then employed to construct features that are more effective for robust and highly accurate image classification via the

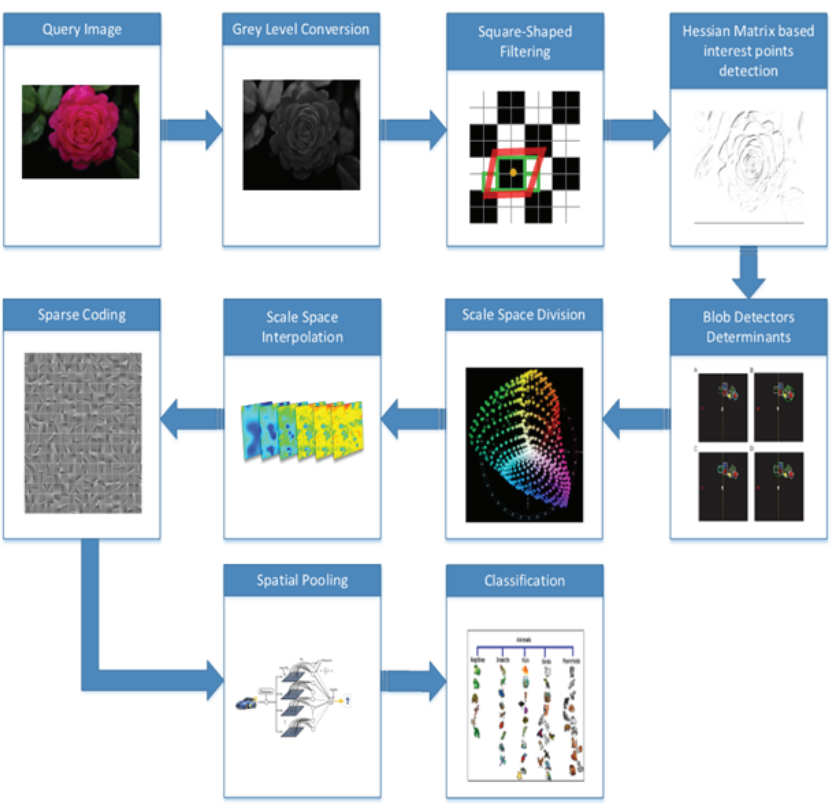

Figure 1: Overview of Proposed Methodology

In addition, the Classification approach demonstrates all the advantages of understanding and computing flexibility. Once the process is set up for classification, the classifier will prioritize the visual elements present in the digital image. Since our method typically employs a linear classifier, its flexibility in preparation scales linearly with the number of training images, in contrast to the quadratic or higher complexity nature experienced by nonlinear bit-based methods.

suggested feature reshaping technique. The SVM performs classification on data that has been prepared by sparse coding and spatial pooling. Training and testing are the two phases of the SVM. Precision refers to the ability to accurately forecast outcomes, while recall refers to the

accuracy with which those outcomes are evaluated. Both small and large databases with 100 images are used to calculate precisions and recalls. In this case, accuracy is achieved by running the proposed method on various subsets of each database on a machine equipped with 4GB of RAM (an HP Pivilion G6 1010tu), demonstrating the algorithm's performance, robustness, and efficiency. Equations (eq. 6) and (eq. 7) show the formulas used to determine the accuracy and recall rates, respectively.

$$\text{Recision} = \text{images (relevant)} + \text{images(retrieved)} / \text{images (retrieved)} \quad (6)$$

$$\text{Recall} = \text{images (relevant)} + \text{images(retrieved)} / (\text{images (relevant)}) \quad (7)$$

3.1. Dataset

Selecting an appropriate picture database for image classification is an important and challenging issue. Each image in the dataset is used as an input image, and similarity between the images is used to determine how each image should be classified. Select photographs at random from the database and put them through their paces in a training and testing phase to ensure accuracy. Results are provided using a mean and standard deviation of recognition rate for each class, with results broken down each run. The Caltech-101 called D1, the FTVL called D2, and the Corel-1000 called D3 datasets are used in the experiments. We have chosen ten classes, each with 100 images, from the D3 collection, all of which are 348 by 256 pixels in size. Images with D2 are 1024x768, and while there are 15 classes available, we've only used 10. Ten of the D1 courses were used, along with 300x200px graphics.

A. All Fruits

The D2 database, shown in Figure 2, has 2,612 images of fruits and vegetables. There are fifteen categories of 1024x768 images in the D2 database. Each of the 10 classes has 100 images and is used to test the suggested algorithm.

Figure 2: Sample of each category (D2 Dataset)

B. Caltech-101

Animals, cars, flowers, the human brain, and human faces are just few of the 102 categories that make up D1. Each image in the database is 300200 pixels in size, and while certain categories may have as many as 800 entries, the vast majority of them typically have between 40 and 50. Airplanes, Things, and Leopard are just a few of the 10 categories with 100 photographs each chosen for testing. The Airplanes class has unique photographs of airplanes, whereas the Brains class contains images of brains. Both the butterfly and leopard groups feature many depictions of the animals. Since there are variations across photographs in the same class, a total of 1000 images were chosen for the image classification. We picked it because of the widespread variation it exhibits across image types. Checking the efficacy of image applications is complicated by the fact that D1 is taught in a variety of formats. Figure 3 displays several example images of D1.

C. Corel-1000

Sample images from each of the 10 categories in the D3 dataset are shown in Figure 4. The dataset as a whole consists of 1000 photographs that are all visually similar. We've chosen 10 categories each with 100 images to classify. The database contains images with a 348-by-256-pixel resolution.

Figure 4: Sample of each category (D3 Dataset)

3.2. Experimental Results

In order to efficiently identify local interest points in images based on intensity, we convert the color space to grayscale. This conversion streamlines image processing and facilitates more effective computation. These points of interest have their

global features retrieved using a sliding window optimization. Feature extraction lessens the workload involved in characterizing a massive dataset. In order to solve the problem of computing cost for classification algorithms, sophisticated data analysis is undertaken. The suggested categorization algorithm outperforms competing algorithms while maintaining an O (n) complexity.

A. Experimental Results of Corel-1000

The effectiveness of the provided method for picture categorization is tested experimentally using industry standard benchmarks. There is a comparison between the proposed method and the status quo. We also make comparisons to the work of [44 - 50] using the D3 dataset. In figure 5, we see a visual depiction of the performance achieved by the proposed method and by other methods now in use while figure 5 Precision (Average): D3 Dataset. Table 1 displays the accuracy rate achieved by the proposed approaches across the various image types in the D3 database. Table 1 displays the comparative accuracy of the proposed approach to that of other algorithms for several categories of the D3 dataset, along with the results of state-of-the-art descriptors and the suggested method. Although their 69% accuracy is lower [49]. it is higher than any of the other methods, including the proposed method. While some competing methods achieve respectable results for three or four categories of the D3 database, the suggested method achieves above-average performance across the board. While the method employed [49]. is more accurate than the others and more similar to our own, the average precision of the suggested method is greater than that of the methods with which it is compared. Since the proposed method achieved a higher degree of precision, it may be concluded that it is more secure, dependable, and efficient than methods achieving a lower degree of accuracy. Table 1 displays comparative results between state-of-the-art techniques and the proposed method, with the later showing more accuracy when compared to the latter for various classes of the D3 dataset. Our current method surpasses the previous methods presented in table 1 over all classes of D3, demonstrating its superior robustness and efficiency.

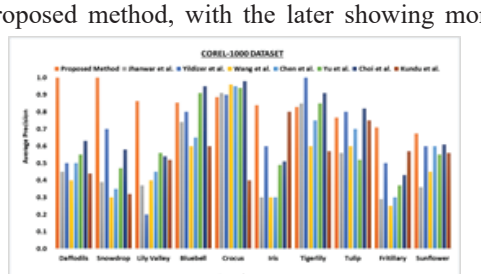

Figure 5: Precision (Average): D3 Dataset

The exceptional performance of the suggested method for the D3 dataset is confirmed by a comparison with other approaches, as shown in table 1. The proposed method achieves a better level of accuracy than the alternatives (78%).

Table 1: Comparison results of Precision (Average) for D3

Proposed Method	T. Wan et al. [29]	M. Balci et al. [30]	Z. Qin et al. [31]	J. Z et al. [32]
0.78 ± 0.09	0.62 ± 0.09	0.66± 0.03	0.48 ± 0.08	0.55 ± 0.05

When compared to the presented method and all other methods, Wang et al.'s 48% precision indicates that their method is less effective in image classification. Table 1 displays the results achieved by the proposed method for image classification tasks, demonstrating its superior efficiency and resilience.

B. Evaluating Against Established Detectors and Descriptors

Visual characteristics of images like texture, color, shape, and region are described by the image descriptors. Because of this, descriptors are crucial in the processes of finding and identifying objects. Many applications rely on popular descriptors like SIFT, HOG, LBP, RGBLBP, DoG, MSER and SURF. Gaussian difference proposed for blob features, salient feature detection, tracking moving objects, denoising medical images, micro classification clusters, and face detection. MSER by Matas detects, recognizes, identifies, and locates lanes and

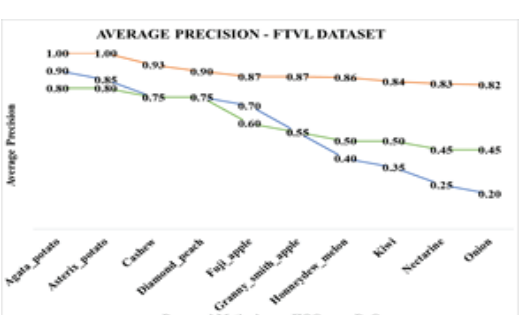

Figure 6: Average Precision of D2

classifies texts. SURF, presented in 2006, used in video stabilization and identifying the retinal optic disc. Proposed method compared to HOG and Difference of Gaussian using four datasets and ten categories each. For the three datasets, we compare the proposed method against establish descriptors using precision (average) graphs.

Figure 6 displays the average precision rates for ten classes extracted from dataset D2, where the suggested technique outperforms DoG and HOG by a factor of 82% to 100%. The outcomes demonstrate the superior effectiveness of the suggested method over competing algorithms. Categories from the dataset D2 are compared to DoG, HOG, and the suggested technique. In comparison to DoG's average accuracy of 0.570 to 0.800, the suggested method achieves results in the range of 0.895 to 1.000, demonstrating the strong robustness of the method for the different classes in the D2 dataset. The average retrieval recall rates for DoG and HOG are 11% to 22% and 12% to 20%, respectively. In comparison to existing methods, the suggested methodology for the D2 dataset performs exceptionally well, with an average retrieval recall rate of between 10% and 11%. Our unique approach is highly effective, robust, and accurate, as evidenced by the reduced recall values we achieved with it. By accomplishing the image classification on the D2 dataset with a lower average retrieval recall and a better retrieval precision rate, the described approach outperforms the other state-of-the-art solution. The experimental charts depicting the performance evaluation of the offered strategy show that the proposed approach has a higher average retrieval recall rate than the other approaches used in the comparison. Table 2 displays the average accuracy of the proposed method, DoG, and HOG for 10 classes drawn from the D2 database: Both the "Onion" and "Cashew" classes benefit greatly from SIFT's superior performance. The proposed technique outperforms MSER for the classes "Agata_potato" and "Asterix_potato," whereas MSER excels for the "Fuji apple" class. Since the lowest average precision of the suggested technique is 82%, whereas HOG (20%), DoG (30%), and SIFT (30%), the presented approach has overall remarkable performance compared to other approaches for all classes of the D2 dataset. When compared to other methods, our method's feature extraction has a far higher average precision rate, proving that it is more resilient, efficient, and successful than the methods to which it is being compared. With spatial pyramid matching and SURF sparse coding, we may extract spatial sparse features that better describe the objects in question when utilizing the support vector machine classifier to categorize images. The proposed method concludes that combining spatial pyramid matching with the SURF sparse coding technique yields performance

Table 2: Precision (Average) for D2

Categories	Proposed	HOG	DoG
Patao(Agata)	100	35	80
Potato (Asterix)	100	40	80
Cashew	86	55	45
Peach (Diamond)	93	75	60
Apple (Fuji)	87	75	75
Apple (Granny Smith)	90	90	75
Melon (Honeydew)	84	70	40
Kiwi	82	85	35
Nectarine	87	25	30
Onion	83	20	55

While HOG and DoG have average precision rates of 10%-95% and 25%-90%, respectively, for chosen classes in the D3 dataset, the suggested technique has an average precision rate of 67%-100%.

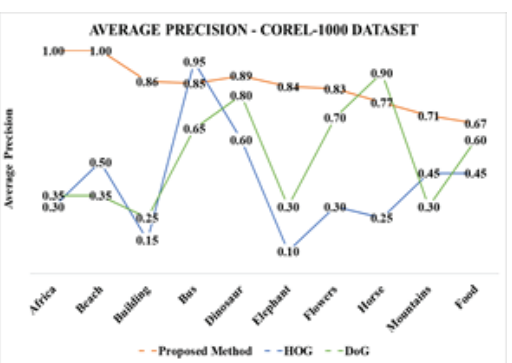

Figure 7: Precision (Average) of D3

As can be seen in Figure 7, the proposed approach has a greater precision rate compared to other algorithms. The suggested method has a higher average precision than competing methods because it makes use of the sparse coding strategy for feature extraction, which improves feature extraction performance. When applied to sparse

image characteristics, maximum spatial pooling improves object detection accuracy, allowing for more accurate image categorization. Digital image object detection could be much improved with the help of spatial information. When compared to other state-of-the-art methods, the suggested method's better accuracy represents its overall remarkable performance. The effectiveness and resilience of the suggested approach are demonstrated by the greater accuracy. Proposed algorithm shows outstanding performance as average retrieval rate is from 80% to 99%, which in higher than DoG and HOG, which represents the higher effectiveness of the presented algorithm as average retrieval rate is from 0.22% to 0.60%. In order to prove the efficacy of fusing SPM and SURF sparse coding, the proposed algorithm is compared to existing state-of-the-art approaches for average precision. HOG has an average retrieval precision of 0.462, DoG of 0.715, and the proposed technique of 0.865. When compared to other methods, the superior performance of the proposed method is indicated by its higher valuer. Table 3 displays the average precision of the proposed method, HOG, and DoG for ten classes from the D3 database. In the "Horse" performance class, DoG excels. While SURF and HOG both perform well for the "Dinosaur" class, the "Bus" class is where HOG really shines, and the "Africa" and "Beach" classes are where the proposed technique really shines. HOG's minimum average precision is only 15%, well below the minimum accuracy of other methods (which is set at 30%). For the D3 dataset, the proposed approach has higher accuracy, robustness, and effectiveness than the alternatives, all of which have minimum performance below 60%. The proposed method's performance, meanwhile, shows average precision greater than 66% across all classes. The mean average recall values achieved by the proposed approach, ranging from 10% to 12%, surpass those of competing algorithms, showcasing its superior performance specifically on classes within dataset D3 when compared to the DoG and HOG methods.

while HOG and DoG have average precision values between 15% and 55% and 20% and 95%, respectively, for selected classes in the D1 dataset, the suggested technique has values between 60% and 100%. The superior accuracy of the suggested method compared to existing algorithms is displayed in Figure 8. The proposed method achieves remarkable success in the

Table 3: Average Precision for D3

Class	Proposed	HOG	DoG
Africa	100	30	35
Beach	100	50	35
Building	86	15	25
Bus	85	95	65
Dinosaur	89	60	80
Elephant	84	10	30
Flowers	83	30	70
Horse	77	25	90
Mountains	71	45	30
Food	67	45	60

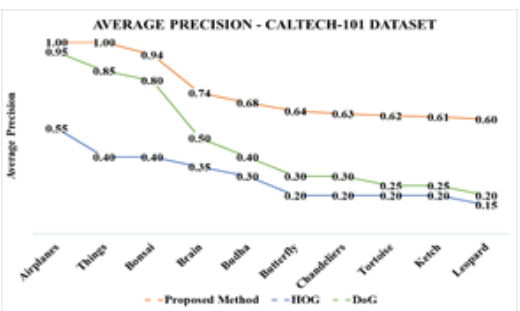

Figure 8: Precision (Average) of D1

"Bonsai" and "Butterfly" categories, while the DoG method comes closer to success but falls short. The blue line in Figure 8 represents the poor performance of HOG on a subset of the D1 database. The proposed method only achieves 60% accuracy for the "Leopard" class, but it does exceptionally well for the other classes, demonstrating its robustness and efficiency. When compared to existing methods, the suggested method for picture classification using spatial pyramid matching with SURF sparse coding yields superior results across all ten classes in the D1 benchmark dataset. Calculating the mean precision for the D1 classes we get the average retrieval accuracy for those objects. DoG yields values between 0.380 and 0.500 HOG, whereas the proposed algorithm yields result between 0.745 and 0.999, demonstrating its superior performance.

The better accuracy of the proposed method indicates the greater stability of the method. The high average retrieval accuracy findings

demonstrate the superiority of the suggested method over previous approaches, which perform with lower average retrieval precision, for chosen classes of D1. The average retrieval recall rates for DoG and HOG are 18% to 31% and 10% to 22%, respectively. In comparison to existing methods, the proposed strategy has a higher average retrieval recall rate of 10% to 14%, demonstrating its superior performance and accuracy. The average retrieval precision and recall for the three benchmark databases are calculated and compared. The resilience and efficacy of the offered approach for picture classification tasks are demonstrated by the fact that the proposed approach achieves greater accuracy across all datasets. Table 4 displays the average precisions of the proposed method, HOG, and DoG for a subset of the classes in the D1 database. The suggested method outperformed the state-of-the-art DoG on the D1 dataset for all but the "Butterfly" class, with an accuracy of 80%. For the D1 dataset, the accuracy of HOG is at a bare minimum 15%, well below that of the other methods compared to which the suggested method is superior.

Compared to competing methods, the one provided here achieves a minimum of 60% accuracy for D1, demonstrating its superior efficacy and resilience. Proposed efficiency is demonstrated by the high degree of accuracy of results across all ten D1 courses.

Table 4: Performance in term of Average Precision for D1

Class	Proposed	HOG	DoG
Air-planes	100	30	50
Things	100	15	20
Bonsai	94	20	20
Brain	74	40	25
Budha	68	15	30
Butterfly	62	20	95
Chandeliers	63	15	30
Tortoise	61	15	25
Ketch	64	35	80
Leopard	60	20	85

4. CONCLUSION AND FUTURE DIRECTIONS

In this research, we have investigated and presented a new method for dealing with broad picture categorization issues. The proposed method for image classification presented in this research is based on SURF sparse codes. When it comes to feature learning, sparse representation is a popular option. Sparse coding is used in this technique to extract superior features from picture block descriptors. To further combine interpretation with scale invariance, sparse coding permits maximal pooling on some spatial measure as opposed to merely relying on average pooling in histogram. The linear-SVM used in this research has been shown to boost testing speed, training flexibility, and classification accuracy. The extract digital image illustration is a prime example of the study's significance. Classification experiments on digital images from different databases demonstrate the effectiveness of the method. While a nonlinear SPM based on vector quantization is widely used in image classification systems, we believe the recommended linear SPM will significantly improve by enabling the application of a more comprehensive collection of trainings. This work suggests that the sparse code of SURF features can serve as a signal for higher local descriptor in common image processing tasks. Some state-of-the-art descriptors and complex algorithms are used to compare experimental outcomes on benchmark databases. We extracted ten categories of images from each database, and each category comprises one hundred images. With most descriptors and many image types, the presented method outperforms the previous approaches.

MATLAB, a high-level programming language useful for completing computationally difficult, is chosen for the implementation in this study. There is a huge library of digital image processing methods in MATLAB, which may greatly simplify and speed up many processes. This research is a method for classifying images based on matching spatial pyramids using sparse SURF codes. This method mixes sparse representation with spatial pyramids to increase speed and accuracy. The approaches employ selective sparse coding as an alternative to traditional vector quantization. This method is utilized to extract valuable features from the appearance descriptors of local image blocks. Additionally, it incorporates scale-invariant local maximum pooling across different spatial scales. The

experimental results clearly showcase the exceptional performance of the proposed method when employed with a linear SVM classifier for the classification task. The three datasets benchmark are used for the experiments. This work suggests that a sparse representation of SURF features can serve as a better local appearance descriptor for typical digital image processing tasks. It would be fascinating to learn more about this through additional experimental study and theoretical consideration. The efficiency of the encoding method is also a concern. On average, it takes a second for each image to go through SURF encoding.

REFERENCE

[1] R. Fergus et al., "Learning object categories from Google's image search," in *Computer Vision, ICCV 2005. Tenth IEEE International,* 2005.

[2] P. Kamavisdar et al., "A survey on image classification approaches and techniques," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 2, pp. 1005-1009, 2013.

[3] A. Koltunov and E. Ben-Dor, "A new approach for spectral feature extraction and for unsupervised classification of hyperspectral data based on the Gaussian mixture model," *Remote Sensing Reviews*, vol. 20, pp. 123-167, 2001.[4] A. Raza et al., "Enhance Voice-Based Email Web System for Visually Impaired People," I*nternational Journal of Recent Technology and Engineering (IJRTE) 11.3 International Journal of Recent Technology and Engineering (IJRTE). Web*, pp. 27–34, 2022.

[5] L. Scrucca et al., "mclust 5: Clustering, classification and density estimation using gaussian finite mixture models," *The R journal,* vol. 8, pp. 289, 2016.

[6] G. Griffin et al., "Caltech-256 object category dataset," 2007.

[7] M. U. Farooq et al., "Bigdata analysis of stack overflow for energy consumption of android framework," In *2019 International Conference on Innovative Computing (ICIC),* (pp. 1-9), IEEE, (November, 2019).

[8] M. U. Farooq et al., "Melta: A method level energy estimation technique for android development," In 2019 *International Conference on Innovative Computing (ICIC),* (pp. 1-10), IEEE, (November, 2019).

[9] W. Liu and E. Y. Wu, "Comparison of non-linear mixture models: sub-pixel classification," Remote Sensing of environment, vol. 94, pp. 145-154, 2005.

[10] C. D. Lloyd et al., "A comparison of texture measures for the per-field classification of Mediterranean land cover," *International Journal of Remote Sensing*, vol. 25, pp. 3943-3965, 2004.

[11] A. Latif et al., "Impact Of Big Data Analytics And Artificial Intelligence On Talent Management".

[12] M. Zortea et al., "A SVM ensemble approach for spectral-contextual classification of optical high spatial resolution imagery," in *Geoscience and Remote Sensing Symposium*, *IGARSS IEEE International,* pp. 1489-1492, 2007.

[13] B. Schölkopf and A. J. Smola, "Learning with kernels: support vector machines, regularization, optimization, and beyond: MIT press," 2002.

[14] S. Lim et al., "Efficient iris recognition through improvement of feature vector and classifier," *ETRI journal*, vol. 23, pp. 61-70, 2001.

[15] S. Khan, et al., "Alignment Finder: An Interactive Ontology Alignment Framework".

[16] M. Waqas et al., "Ensemble-based instance relevance estimation in multiple-instance learning," In *2021 9th European Workshop on Visual Information Processing (EUVIP),* (pp. 1-6), IEEE, (June, 2021).

[17] S. U. R. Khan et al., "GLNET: global–local CNN's-based informed model for detection of breast cancer categories from histopathological slides," *The Journal of Supercomputing*, pp. 1-33. 2023.

[18] A. Raza et al., "Enhancing Breast Cancer Detection through Thermal Imaging and Customized 2D CNN Classifiers," *VFAST Transactions on Software Engineering,* 11(4), pp.

80-92, 2023.

[19] D. A. Clausi, "Comparison and fusion of co-occurrence, Gabor and MRF texture features for classification of SAR sea-ice imagery," *Atmosphere-Ocean,* vol. 39, pp. 183-194, 2001.

[20] S. U. R. Khan et al., “Hybrid-NET: A fusion of DenseNet169 and advanced machine learning classifiers for enhanced brain tumor diagnosis,” *International Journal of Imaging Systems and Technology.*

[21] K. Van De Sande et al., "Evaluating color descriptors for object and scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, pp. 1582-1596, 2010.

[22] C. J. Du and D. W. Sun, "Shape extraction and classification of pizza base using computer vision," *journal of food engineering,* vol. 64, pp. 489-496, 2004.

[23] O. Boiman et al., "In defense of nearest-neighbor based image classification," in *Computer Vision and Pattern Recognition, CVPR, IEEE Conference on,* 2008, pp. 1-8, 2008.

[24] J. Wang et al., "Automatic framework for semi-supervised hyperspectral image classification using self-training with data editing," *in Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2015 7th Workshop on*, pp. 1-4, 2015.

[25] J. Yuan et al., "Factorization-based texture segmentation," *IEEE Transactions on Image Processing*, vol. 24, pp. 3488-3497, 2015.

[26] X. Wang et al., "Fast unsupervised texture segmentation using Texel similarity map," J*ournal of Modern Optics,* vol. 62, pp. 1211-1222, 2015.

[27] J. Yu et al., "Feature integration analysis of bag-of-features model for image retrieval," *Neurocomputing,* vol. 120, pp. 355-364, 2013.[28] Z. Zhang et al., "Noise modeling and representation based classification methods for face recognition," *Neurocomputing*, vol. 148, pp. 420-429, 2015.

[29] E. Yildizer et al.,"Efficient content-based image retrieval using multiple support vector machines ensemble," *Expert Systems with Applications,* vol. 39, pp. 2385-2396, 2012.

[30] C. Wang et al., "Spatial weighting for bag-of-features based image retrieval," in *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making,* pp. 91-100, 2013.

[31] Y. Chen et al., "CLUE: cluster-based retrieval of images by unsupervised learning," *IEEE transactions on Image Processing,* vol. 14, pp. 1187-1201, 2005.

[32] M. K. Kundu et al., "A graph-based relevance feedback mechanism in content-based image retrieval," *Knowledge-Based Systems*, vol. 73, pp. 254-264, 2015.