Single and Cross Domain Image Retrieval using Multi-Modal Feature Fusion

Venkataravana Nayak K, Sharathkumar S K, Arunalatha J S and Venugopal K R, *Fellow, IEEE*

***Abstract*—Image retrieval plays an important role in the analysis of obtained decisive visual information. The presence of visual inconsistency in visual appearance decreases the retrieval accuracy and many of the present retrieval methods emphasize single-source retrieval with the assumption of queries and databases distributions being similar. The number of features obtained with the traditional approach in which some of them are redundant, correlated, and sometimes noisy, increases the model feature space complexity and decreases interpretability. From the study of previous work, it is evident that feature fusion with cross-domain retrieval has not been addressed thoroughly so far. Thus, to deal with these issues, this extracts the optimal combination of the multi-modal features and fuses for enhancing retrieval accuracy. The complementary features obtained are effective with the traditional approach for the improvement of representation and retrieval effectiveness. Thus, Image Retrieval using Single and Cross-Domain Feature Fusion (SCDFF) is proposed in this work. The multi-modal features are extracted with Texture, Color, Statistical, and Scale Invariant Feature Transform (SIFT) descriptors to perform the retrieval process. The feature vector is fused using an optimized weight value which is obtained from Glowworm Swarm Optimization (GSO) algorithm and the image similarity is computed with K-Nearest Neighbor. An empirical analysis is performed to evaluate the proposed model and from the results obtained, it is evident that this work outperforms existing approaches in terms of accuracy. The novelty of this work lies in the fact of Single-Domain Feature Fusion (SDFF) and Cross-Domain Feature Fusion (CDFF) with optimization for Image Retrieval.**

***Index Terms*—Cross-Domain Image Retrieval, Feature Fusion, Glowworm Swarm Optimization, Retrieval Accuracy, Weighted K-Nearest Neighbor.**

I. Introduction

Widespread usage and advancements of capturing devices have led the acquisition and storage of data in datasets easy. The demand for effective image retrieval is increasing based on visual content for scientific research, forensic analysis instantaneous use. Thus, developing operative retrieval models is necessary to meet this demand and the challenge of this is representing dataset images through the efficient features extraction and retrieval of images close to one another.

The multi-modal features of the visual data of images have been generally used to extract minute image details in the pattern recognition area [1], [2]. The feature descriptors find a compact set of salient features in detecting feature values associated with an image that increases recognition and retrieval accuracy [3]. To obtain useful decisive information, the feature sets obtained through the feature extraction process are fused into a single feature set with appropriate feature normalization, transformation, and reduction. This process consents to enhance the quality of retrieval by incorporating spatial and spectral information of samples to increase the applicability of images in a variety of image processing applications.

The representation of the extracted features is significant and it has a close relationship with human perception. The visual search components for perception stimuli express the meaningfulness of scenes or images. The theory of perception directs the formation of individual observable features of an object as complete perception.

In the existing method, the low-level features from the dataset images *i.e.,* Color and Texture features are extracted with the traditional approach, and Deep Features with Convolutional Neural Network(CNN) classifier. The obtained features are optimized with the Genetic algorithm with its weights. The optimal features are fused and selected optimal combination. The closeness of the query image and dataset images is computed with an SVM classifier and based on the closeness values relevant images are retrieved [4].

In the decision-making process of coordinated tasks in which time is crucial, the complexity in terms of processing time has to be reduced. Thus, the deep features part is not used in the proposed method. The combination of features selected is less representative and takes time for the retrieval process. To deal with these drawbacks, the proposed $SCDFF$ method, in which the color features are extracted using the HoG descriptor, texture features from the CSLBP descriptor, feature points from the SIFT descriptor, and statistical features are extracted. The features are optimized with the Glow-worm Swarm algorithm and the closeness of the query image and dataset images is computed with the KNN classifier. The performance of the individual and fused features is measured individually. The proposed method comparatively increases the accuracy and reduces the training and retrieval time [4].

Objective: To enhance the performance of image retrieval by increasing the discrimination ability of image representation with multi-modal features extraction, fusion, and optimization.

Contribution:

- The multi-modal features extraction combination increases features distinctive ability to improve model accuracy by over 10.5%.
- Decreases retrieval time comparatively.
- The Cross-Domain feature fusion approach gives an

Manuscript received May 3, 2022; revised January 12, 2023.

Venkataravana Nayak K is a Research Scholar in the Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, 560001 India (corresponding author e-mail: venkatcsuvce@gmail.com)

Sharathkumar S K is a PG Scholar in the Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, 560001 India (e-mail: sharathkumarsk21@gmail.com)

Dr. Arunalatha J S is a Professor of Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, 560001 India (e-mail: aruna_veeresh@yahoo.co.in.)

Dr. Venugopal K R is Vice Chancellor, Bangalore University, Bengaluru, India (e-mail: venugopalkr@gmail.com.)

accuracy of 95.00%, which is comparatively more by over 9.50%.

Organization: The paper is organized as follows: In section-2, a brief description of visual features and image retrieval is presented. In section-3, an explanation of the background work is presented. The proposed method and algorithm are described in section-4. The experiment analysis is depicted in section-5. The section-6 presents conclusions.

II. Related Works

This section provides brief details about the works related to the single and Cross-Domain feature fusion.

A. Single-Domain Feature Fusion

The features are extracted from imagery data and are fused to find correlated feature values *via* corresponding descriptors towards recognition of compact, distinct features which increases retrieval accuracy.

Chen *et al.,* [5] presented a retrieval framework to analyze the status of various types of the tongue for diagnosis. The color and texture information is obtained with the Generalized Lloyd algorithm and edge-directed histogram descriptor. The distance between the target and source samples is calculated by taking the absolute difference with the addition of weights in the ratio of color to texture is $0.3 : 0.2$. The method uses texture information along with color information to represent tongue image information efficiently. It decreases the time taken in the manual annotation process.

Chu *et al.,* [6] developed an integrated retrieval model for image representation with the theory of perception. The color and texture features are obtained in the HSV color space using the traditional approach and aggregated to accomplish closeness to human perception. Based on the L_1-distance metric relevant samples are retrieved. Results in poor performance due to the use of an insufficient number of features and ineffectiveness in decreasing artifacts.

Zhang *et al.,* [7] devised a framework to retrieve samples of wool based on the color from the color histogram, dominant color and texture features from the color moment descriptor, and color distribution entropy in the RGB color space. Euclidean distance is used to measure the closeness of the samples to retrieve. Reduces time in the analysis of samples and retrieval, robust in differentiating color but gives less performance due to the appearance divergence nature of the sample.

Banharnsakun *et al.,* [8] introduced a retrieval approach based on the optimization of features with an artificial bee colony algorithm. The statistical texture features are obtained with Grey-Level Co-occurrence Matrix descriptor, optimized and the similarity among the samples is computed with Squared Euclidean distance metric. The convergence rate is fast initially and later gradually decreases. The accuracy is less because of inadequacies of single modality features in representation.

Hilasaca *et al.,* [9] designed a feature fusion framework for the retrieval process with low and high-level features. The low-level features such as color from LAB color histogram, texture from Gabor filter and shape features from the HoG descriptor are obtained and the high-level features are obtained with CNN. The features are weighted combined with k-mean clustering and the similarity is computed with the Nearest neighbour classifier. The method shows less scalability.

Liu *et al.,* [10] developed a retrieval model with texture and color features fusion. The texture and color features are extracted with LBP and Color Information Feature (CIF) descriptors respectively. The features are fused and optimized using the Particle Swarm algorithm. The model provides 84.23% retrieval accuracy.

B. Cross-Domain Feature Fusion

The availability of info from any single domain is sometimes not precise, and inconsistent due to its incompleteness. Though the info from multiple sources is useful in obtaining the corresponding or extra information, it is useful in applications that involve feature extraction process and the fusion of cross-domain info helps to obtain consistent and dependable outcomes. Feature extraction from cross domains is a requirement in the applications scenarios of many machine learning applications, for example, the harshness of fire detection in the forest and its land area coverage.

Sahu *et al.,* [11] developed an image search framework based on cross-domain feature fusion. The network parameter features are extracted by the pyshark python wrapper. The log transformation and scaling preprocessing process are performed. Then feature reduction is carried out with the Principle Component Analysis and feature filtering is done with Shapiro ranking and the features are fused. The Elastic-search is used to compute the parameters count. The Gaussian Naive Bayes classifier provides 84% precision, indicating the importance of cross-domain information than a single domain.

Wu *et al.,* [12] introduced the entity retrieval method. It is a Community Question Answering technique to decrease the user activity wait time. Designed a network of information with phrases that are used to represent the relationships among the posts of question and answer and the network of information is fused. The ideal answer is obtained through the entities. The entities are retrieved based on their nearest values by developing a matching algorithm.

Cheng *et al.,* [13] designed a cross-domain feature fusion recommendation framework based on the Minimum Conflict Principle of the Dempster-Shafers theory of evidence. The cross-domain features from the users archive are obtained and computed features relevance using the Backpropagation algorithm. The appropriate closeness of information used in the fusion process can be enhanced towards the improvement of the outcome.

Wu *et al.,* [14] introduced a cross-domain image-matching approach based on edge orientations and keypoints gradients information distribution. Designed an algorithm that identifies the features closeness of cross-domain image pairs. The significant edge features are obtained with the location–oriented boosting detector and histogram of point-edge orientation descriptor for describing the region and decreasing the effects of artifacts. The method eliminates incorrect matches among reference images and dataset samples using a bilateral process of matching.

Shu *et al.,* [15] presented emotion identification approaches using cross-domain information. The communica-

TABLE I: Related Works

Author	Approach	Algorithm	Merits	Demerits	Performance	Applications
Chen *et al*., [2017]	Image Retrieval	Generalized Lloyd algorithm and Edge Histogram extract algorithm	Decreases time taken in the manual annotation process	Accuracy of the model is less	51.13% Accuracy	Medicine
Chu *et al*., [2020]	Image Retrieval	Multi-integration features algorithm, L1-distance metric	Decreases retrieval time	Not effective in decreasing artefacts, Accuracy is less	60.28% Accuracy	Retrieval
Zhang *et al*., [2020]	Image Retrieval	Dominant color and color moments integration algorithm, Euclidean distance metric	Robust in differentiating color	Accuracy of the model is less	87.00% Accuracy	Retrieval of images in the factories fabrication process
Banharnsakun *et al*., [2020]	Image Retrieval	GLCM with artificial bee colony algorithm, Squared Euclidean distance metric	Involves fewer control parameters so fast in convergence at the initial stages	Inadequate in the fulfilment of accuracy requirement	76.00% Accuracy	Advertisement, Online shopping
Hilasaca *et al*., [2020]	Image Retrieval	kmean clustering and Nearest neighbour	Provides better accuracy for the small number of samples	Less scalability	93.65%	Retrieval
Liu *et al*., [2017]	Image Retrieval	LBP, CIF and Particle Swarm algorithm	Provides good performance for the small number of samples	Model scalability is low	84.23%	Retrieval and Classification
Shu *et al*., [2018]	Emotion Identification	Artificial information Integration algorithm	Requirement on priori knowledge of objects is low	Consumes time in information processing, Learning time is more	90.00%	Security, Monitoring and Control, Clinical Medicine etc.,
Wu *et al*., [2020]	Image Matching	Bilateral matching method	Eliminates incorrect matches	Average performance	65.10%	Retrieval and Classification
Cheng *et al*., [2019]	Image Search	Back propagation algorithm	Scalable	Gives poor performance	64.54% Accuracy	E-commerce
Sahu *et al*., [2020]	Detection	PCA, Shapiro ranking, Gaussian Naive Bayes classifier	Involves processing complexity	Provides good performance	84.00% Accuracy	Security and Monitoring

tion of a variety of information *i.e.,* sight and hearing, image, text, speech, physique pose, and physiological indications are collectively considered in getting dependable decisions. The data-level integration consumes time for processing information in the case of large-size data and information loss happens in the case of character-level integration and due to this, the precision obtained at the decision level is less. The cross-domain information has to be optimized appropriately to increase the emotion status recognition precision.

Through the related works, it is observed that the combination of features selected is less representative, and takes time for the retrieval process. To deal with these drawbacks, the proposed $SCDFF$ method, in which the low-level features are extracted, fused, and optimized with the Glow-worm Swarm algorithm then the closeness of the query image and dataset images is computed with the KNN classifier to obtain a stable result. Comparatively, the proposed $SCDFF$ method guarantee accuracy.

III. Background and Motivation

In the existing approach, the low-level features of dataset images such as color, and texture are extracted with the traditional approach, and the deep features are obtained through the CNN classifier. The obtained low-level and deep features are optimized with a Genetic algorithm with its weights. The optimized features are fused and the best combination of features is chosen with a partial selection method. The distance between the input image vector and the database images vector is computed with an SVM classifier. Based on the similarity value the relevant images are retrieved and the performance of the process is measured.

In the feature extraction phase of the existing approach, the combination of the features used is obtained with low-level features of color and texture are Haar, Color, and SIFT descriptors.

- Haar features: The features of images are obtained with the Haar descriptor. The image is segmented into rectangular regions, in each region, the intensity values are combined and subtracted the values from the regions of the image, then combined from the features and trained with the AdaBoost classifier to make them faster. This process is complex and consumes time.
- Color features: The basic red, green, and blue components are obtained with the color descriptor. The different types of variations are not taken into account.
- SIFT features: SIFT descriptor extracts fever, large size, and tuned parameter but the rotation, and scaling factors are not considered.

The features combination used in the features extraction phase of the existing retrieval process is not comparatively optimal. Due to the exclusion of geometric and photometric variations, the structural information obtained is not adequate. The features obtained are optimized with a Genetic algorithm with its weights. The result obtained is suboptimal, as the iteration count increments the interdependency of the suboptimal result relies on the initial size of the samples. The features are derived from the sample *i.e.,* derived features, but the algorithm suites more for the direct input with respect to giving consistent results. Which impacts processing speed [16]. Hence, instead used glowworm swarm optimization, as the iteration count increments the vision range of glowworms decreases, it split up glowworms which involves searching neighbors appropriately and helps in excluding false positives.

The case of the retrieval process with deep features involves processing time. The complexity involved in the decision making process has to be reduced in coordinating tasks of robot applications, where time and memory resources are critical constraints. The local features are fast and adaptable; with these benefits, we are using the traditional approach of feature extraction.

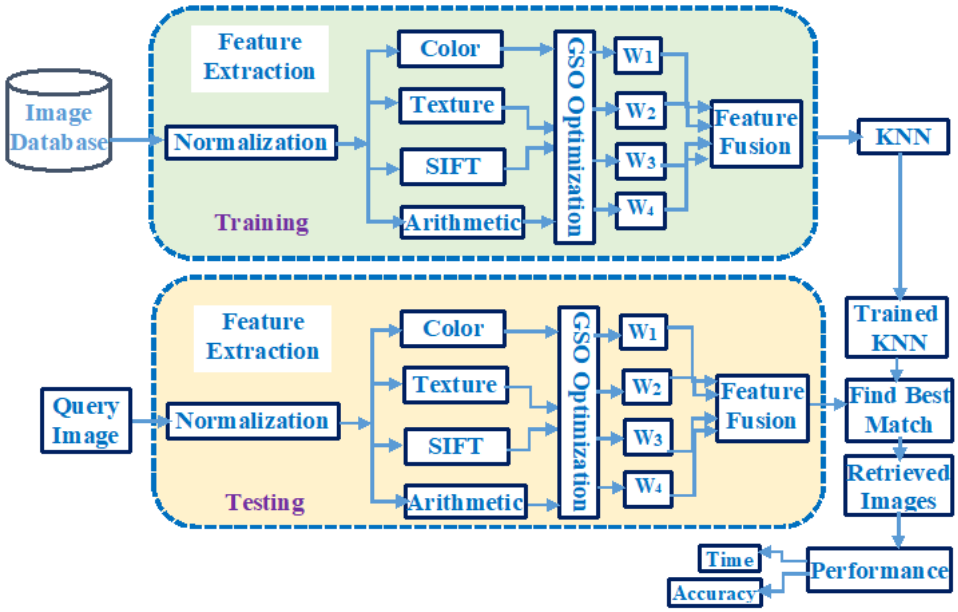

Fig. 1: The Proposed System Architecture ($SCDFF$)

The SVM classifier is used for training and similarity computation. In case of an increased number of features than the number of samples, it will not give consistent results. Hence, it is not appropriate for the dataset to have more samples. Due to this reason, a kNN classifier is used. It learns during the testing phase by having training information in memory and in case of an increased number of samples, it gives stable results.

There is no single method sufficient in obtaining all types of features of the images [17]. The design of each of the detectors and descriptors are individually strong in certain types of features respectively. The local features are fast and adaptable [18], [19], and this advantage is used in the feature extraction phase. One way to make a combination of obtained features optimal is by using the strength of local features, to reduce complexity in the decision-making process of cooperating tasks by taking care of geometrical and photometric variations and memory storage constraints.

In the proposed approach, the low-level features of dataset images such as color, and texture are extracted with the traditional approach. The obtained features are optimized with the Glowworm swarm algorithm with its weights. The optimized features are fused and stored in a vector. The vector features are trained with the kNN classifier and the distances between the input feature vector and the trained feature vector are computed. Based on the similarity value the relevant images are retrieved and the process performance is evaluated.

IV. Proposed Method

A. Problem Statement

Develop a retrieval model using visual feature fusion and optimization for increasing features discriminative ability to obtain a stable result.

Objectives:

(i) To increase the retrieval accuracy
(ii) To reduce retrieval time and
(iii) To reduce the training time

The proposed architecture is shown in Fig. 1. The dataset images are resized and normalized, the color, texture, and arithmetic features are obtained with the traditional approach. The obtained features are optimized with a glowworm swarm algorithm with its weights. The optimized features are fused and stored in the vector. Using kNN, the optimized features are trained and the similarity distance between the query vector and dataset vector is computed. Based on the similarity value, relevant images are retrieved and the performance of the process is evaluated.

B. Feature Extraction

In the feature extraction phase, the images are resized to $32 * 32$, 64-bit double-precision format is used to reduce the roundoff error of pixels and to make extracted features combination comparatively optimal, we have obtained Color (HoG), Texture (CSLBP, SIFT) and Arithmetic features with the traditional approach.

- **HoG:** The magnitude and direction features are obtained to better differentiate the images. To reduce the length of the feature vector, the histogram bin size is taken as 20.
- **CSLBP (Centre Symmetric Local Binary Pattern):** Compact version of LBP, in an eight neighborhood, it takes four comparisons to compute the pixel value, so it reduces the length of the feature vector. Contribute to the reduction of similarity matching time.
- **SIFT:** Initially, it chooses the feature point in the image space, then it gives location, scale, and orientation.
- **Arithmetic features:** The fundamental information of the data is obtained, represents corner positions, reduces noise and variations.

C. Glowworm Swarm Optimization (GSO)

A computational approach [20], in search space, optimizes the glow-worm population iteratively with respect to its

location and movement. The location of the Glow-worm locally best influences each one movement and it is updated as a better location followed by other glow-worms, it moves towards the best solution.

1) Phases of GSO: A Glowworm Swarm GS contains G Glowworms, initially in the solution space distributed randomly [20]. Each of $G_i(i = 1, 2....n)$ has random position RP_i, local decision range LDR_i and luminescence component level LL_i parameters. LL_i relay on RP_i and $F(t)$.

- **Initialization:** The parameters ρ, $gamma$, $beta$, CSR, $Location$, *etc.,* are critical, and impact the outcome. All G_i randomly occupy locations in the solution space, which consists same quantity of luminescence, CSR values, and current iteration $t = 0$.
- **Luminescence updating:** Relays on the function value at the current G_i locations. In every iteration a LL_i is updated based on the G_i changed location as in Equation 1.

$$LL_i(t) = (1 - \rho)LL_i(t - 1) + \gamma F(RP_i(t)), \quad (1)$$

A fraction of Luminescent value ρ is subtracted from it to motivate decay with time. where, t - current iteration, $LL_i(t - 1)$ - previous Luminescent level for G_i, ρ - Luminescent decay constant- $\rho \in (0, 1)$ and γ- Luciferin enhancement fraction, $F(RP_i(t))$- objective function value for G_i at current position RP_i.

- **Neighbourhood selection:** The G_i is selected with a probabilistic approach based on LL_i values. The G_i having higher LL_i attracts G_j to move towards it. The G_j will be selected as neighbors if they are located within the local decision domain of G_i and are given by Equation 2.

$$SG \in NS_i(t)\ iff\ WED_{ij} < LD_i(t)\ and\ L_j(t) > L_i(t) \quad (2)$$

- **Moment probability computation:** Within a variable local decision domain, each G_i attracts brighter G_j. The G_i seek a neighbor with a probabilistic approach having higher LL_i and move toward it. Through the probability of all neighbours, the best neighbor is selected from the neighbor set based on the probability of each G_i moving toward a neighbour G_j is given by Equation 3.

$$MPij = \frac{LL_j(t) - LL_i(t)}{\sum_{k \in NS_i(t)} LL_k(t) - LL_i(t)} \quad (3)$$

- **Movement direction updation:** Every G_i chooses its direction *via* the Roulette Wheel rule. From the $NS_i(t)$, the G_i having higher probability has more chance of selecting as NP_j neighbor position, and the position of selected G_i *i.e.,* GP_i is adjusted by Equation 4.
  The distance between G_i and G_j at time t is WED_i, $j(t)$. Then G_i movement is given by Equation 4

$$GP_i(t) = \frac{GP_i(t - 1) + cGP_j(t) - GP_i(t)}{WED_{ij}} \quad (4)$$

  where, $c-$ is a constant.
- **Updating local decision domain:** It is a dynamic value *i.e.,* a function represents a number of peaks captured.

TABLE II: Notations used

Notations	Meaning
$G_i(i = 1, 2....n)$	Number of Glowworms
GS	Glowworm Swarm
RP_i	random position
LDR_i	local decision range
LL_i	luminescene component level
ρ	Luminescent decay constant
γ	Luciferin enhancement fraction
WED	Weighted Euclidean Distance
NS	Nieghbors Selection
SG	Glowworm Selection
$MPij$	Movement Probability
GP	Glowworm position
NP	neighbour position
CSR	Circular Sensor Range

The LDR_i of each G_i is adaptively updated with Equation 5.

$$LDR_i(t) = min\{CSR, max[0, LDR_i(t - 1) + \alpha(nt - |NP_i(t - 1)|)]\} \quad (5)$$

$CSR-$ Circular Sensor Range, $LDR_i(t - 1)$ previous LDR_i, $nt-$ is a parameter for restricting $NS_i(t)$ size, $\alpha-$ the Model constant. In constructing model, without considering the LDR update step, nt and α, are considered with the same values of LDR_i and CSR. The notations used are shown in Table II.

Algorithm 1 The steps in GSO optimization process

Input: Dataset Images Features
Output: Optimized Features

1: Generation of a Glowworm Swarm GS *i.e.,* G_i, where, $i = 1, 2....n$ and Initialization of parameters: luminescent value-LL_0, luminescent elimination coefficient-ρ, luminescent update coefficient-γ, local domain update coefficients- β, c, iteration number-t, local decision range-LDR_0, circular searching radius-CSR
2: Update Luminescence level: In every iteration, LL_i updated based on the G_i changed location using Equation (1)
3: Selection of Neighbors: The G_i are selected with the probabilistic approach based on LL_i values using Equation (2)
4: Computation of Moment probability: In the variable local decision domain j is chosen by Pij, G_i movement toward G_j is calculated using the probabilistic approach with Equation (3)
5: Movement direction updation: Every G_i choose location *via* Roulette Wheel rule using Equation (4)
6: Local Decision Domain updation: LDR_i of each G_i is adaptively updated using Equation (5)

2) Psuedocode of GSO: The strength of GSO is increased to obtain a better accuracy rate by considering the following: isolated neighbor, step count variation, and use of weights in computing Euclidean distance.

- Without neighbors: The distribution of glowworms is random in nature in the beginning, so in case of the existence of more number candidate solutions creates the possibility of a few individuals not finding neighbors. If individual G_i has no neighbors then it moves randomly

at one step in the solution space of its own decision domain in order to avoid slow convergence due to its stagnant nature.

- Step size: Each glowworm moves with a fixed step length. If it is large, it decreases the rate of convergence, then there is a provision of crossing over the optimal solution *i.e.,* the probability of missing the optimal solution is more. If it is small, and precipitately falls into the local optimal solution, the probability of obtaining an optimal solution is less *i.e.,* decreases performance, becomes low. As the number of iteration increments, the glowworms movements of step size has to change accordingly for obtaining better performance. Thus, in the beginning, a higher value has to be maintained, it helps to avoid the probability of jumping the optimal solution point and with iteration increment, its value decreases to a minimum fixed value; it helps to increase the rate of convergence at later stages to get optimal performance.
- Distance between glowworms: In GSO, we have used the weighted Euclidean distance measure to compute the distance among the input and dataset vectors.

D. Weighted K-Nearest Neighbor

The KNN takes unlabelled images as input and gives an appropriate result through learning essential structures from the labeled input datasets [21]. It works with the assumption of similar points present are closer to one another most of the time to compute similarity. The best value of k *i.e.,* the number of nearest neighbors for the given input is selected based on the value that gives an accurate outcome with the capability of the algorithm with a lesser error rate while running the algorithm. Instead of processing the training record and learning a model, kNN stores them, processing, and learning takes place when it gets a test sample, it uses the stored record from the memory in order to find the class it belongs to during classification.

The KNN classifier requires the following: a set of stored records, a Distance metric to compute the distance between records, the value of k, and the number of nearest neighbors to retrieve.

KNN works in two phases: (a) in the Training phase: saves the records in the data structure to enable searching faster. (b) in the Validation phase: get the test input, and finds k training samples.

In high dimensional data, the Euclidean distance for the two dataset points (PT1, PT2) *i.e.,* 111111111110, 011111111111, binary data depicted gives distance value 1.14142 *i.e.,* intuitive outcome. For another set of points pair (PT3, PT4) *i.e.,* binary data also gives the same value as 1.14142. It basically computes the distance between points without the knowledge of the distribution of data points. To check the uniformity of the distribution of information, the difference between pixels in the neighborhood is considered in the feature extraction phase of the proposed method for reducing the feature vector length.

To overcome the above drawbacks, two approaches are used (a) weighted distance metric and (b) Large value for k.

(a) weighted distance metric: if the feature has a larger weight, it is more important, if the feature has a smaller weight, it is less important, and the feature having zero weight does not matter. The weight value is decided and fixed based on importance, range, and scale features so that they have a similar range or normalize so that they have the same mean and standard deviation. If a feature has a larger range, use small weights and if the feature has a smaller range, use larger weights. Weights allow kNN to be effective with axis parallel elliptical classes.

(b) Large value for k: In the case of $1-NN$, the decision boundary is not smooth. With the small value of k, if there exist fine structures in the problem space will be captured and may be necessary for the small training dataset. If the value of k is large, the neighbors are classified appropriately, the classes are smooth. The large value is appropriate in the following cases, if the classifier is less sensitive to noise, get a better probability estimate for the discrete classes and a larger dataset allows the use of larger k. In the training set for the increased k value, in the midway, we see the best value.

If an image space is defined in terms of a large number of feature attributes, it stances a problem in describing an appropriate similarity metric and for various types of learning problems due to the feature importance or irrelevancy from one another. Precisely it impacts the kNN algorithm critically. So reducing extra features is important. Because in a very high dimensional feature space, two items may be similar but still a difference in unimportant features and differences in distances between different pairs of items are almost similar. So it makes it difficult to find good representative training features for a given test input.

The prediction is based on the weighted average in weighted kNN, the weight is based on the distance *i.e.,* the difference in the distance between two items. The locally weighted averaging is another type of distance; it gives flexibility in choosing a very large value for k. Allows assigning different weights to different training features. The weights fall–off rapidly with distance. In this case, we can choose more kernel width if the neighborhood area is large, we can consider more area in a neighborhood. The width of the kernel controls the size of the neighborhood which has a large effect on value (similar to k). In the case of $k > 1$, choices the possibility of giving different weights to k-nearest neighbors because of distance. Allows computing closeness and closest points in a faster way.

Algorithm 2 The weighted kNN strategy

Input: Query Image, Dataset records
Output: k-training items

1: Compute weighted Euclidean distance between query and training records $X_p = x_{p1}, x_{p2}, x_{p3}....x_{pN}$ and $X_q = x_{q1}, x_{q2}, x_{p3}....x_{qN}$ $D(X_p, X_q) = \sqrt{\sum_{r=1}^{N} w_r.(X_{pr} - X_{qr})^2}$ where, $w_r = \frac{1}{D(C_d, C_{query})}$
2: Find k nearest neighbors
3: Determine new record class label with nearest neighbors class labels *i.e.,* by Locally Weighted Averaging $prediction_{test} = \frac{\sum_{r=1}^{k} w_r * value_i}{\sum_{r=1}^{k} w_r}$ where, $w_k = \frac{1}{e^{KernelWidth * D(C_d, C_{query})}}$

V. Results and Discussions

The proposed $SCDFF$ method is simulated in MATLAB in an $i7$ processor with 8 GB RAM and tested on three datasets. The performance related to the accuracy and retrieval time is measured and compared with the existing methods.

A. Datasets

Three datasets namely Wang, Oxford Flowers, and ImageNet are used to validate the proposed method, each of the datasets has colored photographs of various objects and is classified into different classes.

- Wang: It has 10 classes each of the class has 100 images, in total it has 1000 images [22] with a dimension of 384*256 among 1000 images 900 are used in training and 100 are for testing.
- Oxford Flower: It contains images of flowers that are commonly present in the U.K. It has 17 classes and each class has 80 images, a total of 1360 images [23] among 1360 images, 1020 are used in training and 340 for testing.
- ImageNet: It contains 14,197,122 images [24] used in the classification and retrieval of multi-class images. We have considered 3 classes, a total of 8000 images with the dimension of $384*256$; 6000 are used in the process of training and 2000 for testing.

The sample images of the Wang, Oxford Flowers, and ImageNet datasets are shown in Fig. 2, Fig. 3, and Fig. 4 respectively.

Fig. 2: Sample images of Wang Dataset

B. Performance Analysis

The performance of the proposed method is analyzed with Single and Cross-Domain feature fusion approaches.

1) Single-Domain Feature Fusion: In this section, the training time, and retrieval accuracy details of Single-Domain feature fusion are discussed.

(a) Training Time: The resulted in training time values for the ImageNet, Wang, and Oxford Flowers databases are tabulated in Table III, Table IV, and Table V. The tables depict the time required for training individual features and the fusion of features. The discriminative ability of features increases with the fusion of features; this makes the process of image matching easy. Thus, the fusion approach takes less time to train and retrieve images than the individual features approach.

(b) Retrieval Accuracy: Accuracy is the fraction of the number of relevant items retrieved to the non-relevant items. It is the closeness of measured value with respect

Fig. 3: Sample images of Oxford Flower Dataset

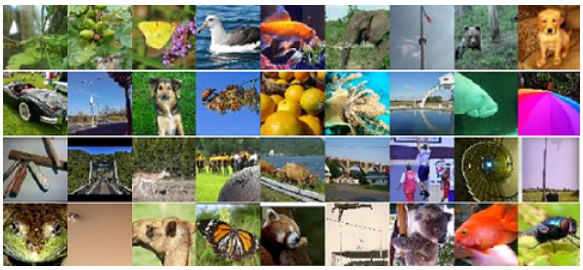

Fig. 4: Sample images of ImageNet Dataset

to the known/standard value and computed as follows: $Accuracy(\%) = \frac{No.\ of\ relevent\ images\ retrieved}{No.\ of\ retrieved\ images} * 100;$

$$= \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

where, the parameters *TP*, *TN*, *FP*, and *FN* represents true and false positives and negatives in the evaluating metric. For the image retrieval model, the performance measures are considered with the limitation of giving *"1"* as the actual value to the retrieved relevant image and *"0"* for the retrieved non-relevant image. Thus the values of *TN* and *FN* become *zero*.

The accuracy relies upon the obtained number of features of the input image. From the dataset images, a total of four sets *i.e.,* HoG, SIFT, CSLBP, and Statistical features

TABLE III: Training Time (Seconds) of different number of ImageNet Database

Image Features	1000	3000	6000
Color (HoG)	022.4227	068.4019	0160.2432
SIFT	341.0925	620.9350	1528.1777
Arithmetic	044.3314	105.8116	0216.6280
Texture (CSLBP)	093.4074	247.5652	0414.1776
Without Fusion	**501.2540**	**1042.7102**	**2319.2213**
SDFF(Proposed)	**374.7920**	**0812.5072**	**2033.7000**

TABLE IV: Training Time (Seconds) of different number of Wang Database

Image Features	1000
Color (HoG)	031.6984
SIFT	391.0657
Arithmetic	040.7306
Texture (CSLBP)	088.5687
Without Fusion	**552.0600**
SDFF(Proposed)	**466.0802**

TABLE V: Training Time (Seconds) of different number of Oxford Flower Database

Image Features	1360
Color (HoG)	034.9449
SIFT	467.1700
Arithmetic	043.4592
Texture (CSLBP)	114.9056
Without Fusion	**660.4797**
SDFF(Proposed)	**542.4604**

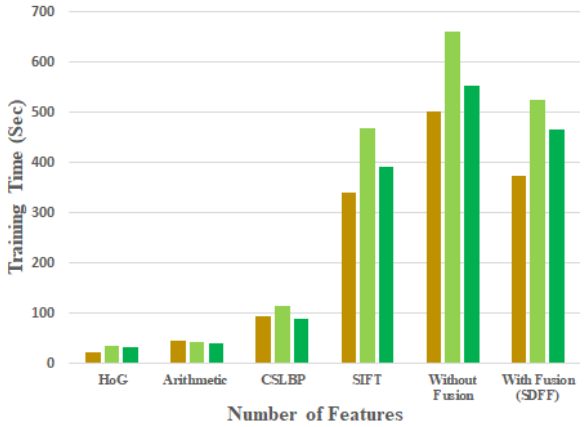

Fig. 5: Training Time of ImageNet, Oxford and Wang Dataset

are extracted. Initially, the system is tested for individual features. Finally, the system is tested with all the features, overall the system produces comparatively more accuracy when the obtained number of features is fused. For the ImageNet Dataset, the number of retrieved images is 100, out of which 4 are unrelevant *i.e* false positives so the result obtained is 96%, tabulated in Table VI, and is shown in Fig. 6.

Similarly, the retrieval accuracy values with respect to the individual features and fusion features for the Wang, and Oxford Flowers are tabulated in Table VII, and Table VIII. The tabulated values show that the retrieval accuracy of the proposed Single-Domain method is higher than the traditional method and takes far less time and does not need GPU. The retrieval accuracy of the proposed model with respect to the individual and fused features is shown in Fig. 7. It shows that the performance of the proposed method is better compared to the conventional technique [22]. The existing Feature Fusion with Genetic Algorithm approach provides an average of 85.50% retrieval accuracy [4]. The closeness of the two visuals for the training samples gives a stable result and enhances unified decision making [26]. The proposed approach gives an average of 96.00%, 97.00%, and 96.50% retrieval accuracy for ImageNet, Wang, and Oxford Flowers datasets respectively. **(c) Comparison of Retrieval Accuracy:** The retrieval accuracy values for various datasets are tabulated in Table VI, Table VII, and Table VIII. The

TABLE VI: Retrieval Accuracy (%) on ImageNet Dataset

Image Features	1000	3000	6000
Color (HoG)	76.20	78.90	79.90
SIFT	85.30	87.40	89.50
Arithmetic	89.60	91.20	92.80
Texture (CSLBP)	92.10	93.40	94.80
SDFF(Proposed)	**96.00**	**96.00**	**96.00**

TABLE VII: Retrieval Accuracy (%) on Wang Dataset

Image Features	1000
Color (HoG)	80.00
SIFT	89.50
Arithmetic	93.00
Texture (CSLBP)	96.00
SDFF(Proposed)	**97.00**

TABLE VIII: Retrieval Accuracy (%) on Oxford Flower Dataset

Image Features	1360
Color (HoG)	80.00
SIFT	89.00
Arithmetic	92.00
Texture (CSLBP)	96.00
SDFF(Proposed)	**96.50**

proposed $SDFF$ method produces accuracy higher than the existing approaches [4], [7], [8], and [10] comparatively and is depicted in Fig.8.

2) Cross-Domain Feature Fusion: In this section, the details about training time, and retrieval accuracy of Cross-Domain feature fusion are briefed. **(a) Datasets:**

- Yale Face Dataset: It consists of 165 images in GIF format in 15 classes, each class has 11 images [25]. The sample images are shown in Fig.9.

(b) Training Time: The samples of the ImageNet and Yale dataset are trained and the resulted in training time values are tabulated in Table IX shows that the time required for training Cross-Domain feature fusion is more than the Single-Domain fusion due to an increase in the size of the samples. **(c) Retrieval Accuracy:** The Cross-Domain feature fusion produces an accuracy of 95.00% tabulated in Table IX shows that it produces accuracy near the Single-Domain feature fusion approach.

(d) Comparison of Retrieval Accuracy: The proposed Cross-Domain model retrieval accuracy value is compared to the methods [14] and [15] are tabulated in Table X shows that the performance is superior to the compared methods and is depicted in Fig.10.

VI. Conclusion

A Single and Cross-Domain Image Retrieval model using the Feature Fusion strategy(SCDFF) is developed and tested using ImageNet, Wang, and Oxford Flowers datasets. The Single Domain multi-modal features fusion approach provides 96.00%, 97.00%, and 96.50% retrieval accuracy respectively and it is comparatively better than 85.50% for the ImageNet dataset carried out by [4]. The model

TABLE IX: Training Time (Seconds) and Retrieval Accuracy(%) of Cross-Domain feature fusion

Image Features (1000)	Training Time	Accuracy
SDFF(Single-Domain)	**374.7920**	**96.00**
CDFF(Cross-Domain)	**386.0042**	**95.00**

TABLE X: Comparison of Retrieval Accuracy (%)

Author	Accuracy(%)
Wu *et al.,* [14]	65.10
Shu *et al.,* [15]	90.00
CDFF(Proposed)	**95.00**

Fig. 6: Retrieval Accuracy of ImageNet

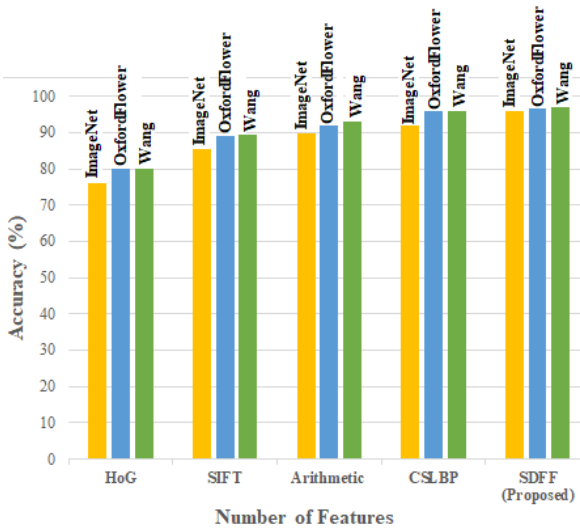

Fig. 7: Retrieval Accuracy (SDFF) of ImageNet, Oxford and Wang Datasets

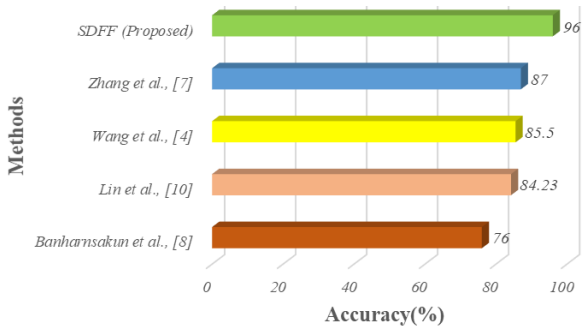

Fig. 8: Comparison of Retrieval Accuracy(SDFF) with Existing Methods

is extended to the Cross-Domain features fusion, where it gives an accuracy of 95.00% and it is near to the Single Domain feature fusion approach. The optimal combination of features obtained increases the features discriminating ability in the image representation. Useful in applications with crucial requirements is fast image retrieval. The model scalability and performance can be increased further using deep features and its optimization. The future work focus is to develop an efficient domain adaptive image retrieval model through a feature fusion approach with big data and analyze deep features behavior and patterns to increase the model interpretability.

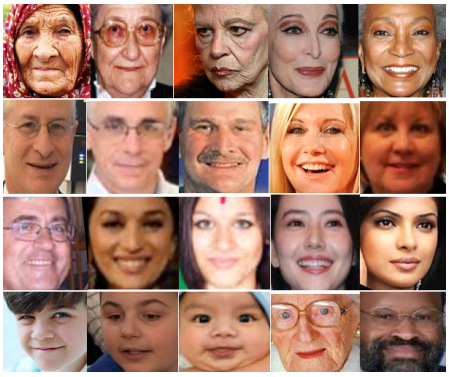

Fig. 9: Sample images of Yale Dataset

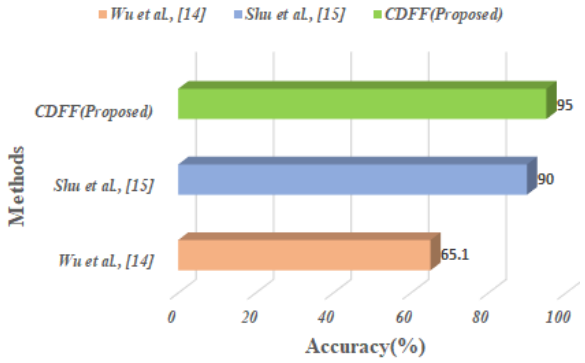

Fig. 10: Comparison of Cross Domain Retrieval Accuracy

ACKNOWLEDGMENT

I gratefully acknowledge the assistance of my supervisor Prof. Arunalatha J S. I would like to thank Prof. Venugopal K R for his encouragement and unwavering guidance.

REFERENCES

[1] Y. Luo, Y. Wen, D. Tao, J. Gui, and C. Xu, "Large Margin Multi-Modal Multi-Task Feature Extraction for Image Classification," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 414-427, January 2016.

[2] X. Lu, L. Song, R. Xie, X. Yang, and W. Zhang, "Deep Binary Representation for Efficient Image Retrieval," *Journal of Advances in Multimedia*, vol. 17, no. 1, pp. 1-11, November 2017.

[3] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous Feature Learning and Hash Coding with Deep Neural Networks," *Proc.of the IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3270-3278, June 2015.

[4] Y. Wang, B. Song, P. Zhang, N. Xin, and G. Cao, "A Fast Feature Fusion Algorithm in Image Classification for Cyber Physical Systems," *IEEE Access*, vol. 5, no. 1, pp. 9089-9098, May 2017.

[5] L. Chen, B. Wang, Z. Zhang, F. Lin, and Y. Ma, "Research on Techniques of Multifeatures Extraction for Tongue Image and Its Application in Retrieval," *Computational and Mathematical Methods in Medicine*, vol. 17, no. 1, pp. 1-11, March 2017.

[6] K. Chu and G. H. Liu, "Image Retrieval Based on a Multi-Integration Features Model," *Mathematical Problems in Engineering*, vol. 20, no. 1, pp. 1-10, March 2020.

[7] N. Zhang, J. Xiang, L. Wang, N. Xiong, W. Gao, and R. Pan, "Image Retrieval of Wool Fabric. Part II: based on Low-level Color Features," *Textile Research Journal*, vol. 90, no. 7, pp. 797-808, April 2020.

[8] A. Banharnsakun, "Artificial Bee Colony Algorithm for Content-based Image Retrieval," *Computational Intelligence*, vol. 36, no. 1, pp. 351-367, February 2020.

[9] G. M. Hilasaca and F. V. Paulovich, "Visual Feature Fusion and its Application to Support Unsupervised Clustering Tasks," *Information Visualization*, vol. 19, no. 2, pp. 163-179, April 2020.

[10] P. Liu and J. M. Guo, "Fusion of Color Histogram and LBP-based Features for Texture Image Retrieval and Classification," *Information Sciences*, vol. 390, no. 2, pp. 95-111, June 2017.

[11] A. Sahu, Z. Mao, P. Wlazlo, H. Huang, K. Davis, A. Goulart, and S. Zonouz, "Multi-source Data Fusion for Cyberattack Detection in Power Systems," *arXiv preprint*, vol. 79, no. 7, pp. 1-18, January 2021.

[12] Y. Wu, S. Zhao, and R. Guo, "A Novel Community Answer Matching Approach based on Phrase Fusion Heterogeneous Information Network," *Information Processing and Management*, vol. 58, no. 1, pp. 1-22, January 2021.

[13] S. Cheng, B. Zhang, G. Zou, M. Huang, and Z. Zhang, "Friend Recommendation in Social Networks based on Multi-source Information Fusion," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 5, pp. 1003-1024, May 2019.

[14] Q. Wu, G. Xu, Y. Cheng, W. Dong, L. Ma, and Z. Li, "Histogram of Maximal Point-edge Orientation for Multi-source Image Matching," *International Journal of Remote Sensing*, vol. 41, no. 14, pp. 5166-5185, July 2020.

[15] Y. Shu and H. Zhang, "Multimodal Information Fusion based Human Movement Recognition," *Multimedia Tools and Applications*, vol. 79, no. 7, pp. 5043-5052, February 2020.

[16] A. S. Girsang and D. Tanjung, "Fast Genetic Algorithm for Long Short-Term Memory Optimization," *Engineering Letters*, vol. 30, no. 2, pp. 528-536, 2022.

[17] S. Krig, "Local Feature Design Concepts," *Journal of Computer Vision Metrics*, vol. 1, pp. 115-166, September 2016.

[18] Y. Liu, H. Zhang, H. Guo, and N. N Xiong, "A Fast-brisk Feature Detector with Depth Fnformation," *Journal of Sensors*, vol. 18, no. 11, pp. 3908-3916, November 2018.

[19] M. Oszust, "An Optimisation Approach to the Design of a Fast, Compact and Distinctive Binary Descriptor," *Signal, Image and Video Processing*, vol. 10, no. 8, pp. 1401-1408, November 2016.

[20] Y. Chen, S. Wang, W. Han, Y. Xiong, W. Wang, and L. Tong, "A New Air Pollution Source Identification Method based on Remotely Sensed Aerosol and Improved Glowworm Swarm Optimization," *Signal, Image and Video Processing*, vol. 10, no. 8, pp. 3454-3464, April 2017.

[21] R. I. Chang, S. Y. Lin, J. M. Ho, C. W. Fann, and Y. C. Wang, "A Novel Content based Image Retrieval System using k-means/knn with Feature Extraction," *Computer Science and Information Systems*, vol. 9, no. 4, pp. 1645-1661, April 2012.

[22] Wang-Dataset https://www.kaggle.com/ambarish/wangdataset.

[23] Oxfordflowr-Dataset https://www.kaggle.com/cantonioupao/oxfordflower-17categories labelled.

[24] ImageNet -Dataset https://www.kaggle.com/c/imagenetobject-localization challenge/overview/description.

[25] Yale-Dataset http://vision.ucsd.edu/datasets/yale-facedataset original/yalefaces.zip

[26] Zhang, Chunjie and Cheng, Jian and Tian, Qi, "Multiview Semantic Representation for Visual Recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 5, pp. 2038-2049, 2018.

Mr. Venkataravana Nayak K is a Research Scholar in the Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, India. He received the Bachelor of Engineering degree in Computer Science and Engineering from Visvesvaraya Technological University in the year 2003 and the Masters of Engineering degree in Computer Science and Engineering from Bangalore University in the year 2006. His research areas of interest are Digital Image Processing and Artificial Intelligence.

Mr. Sharathkumar S K is a postgraduate student in the Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, India. He received his Bachelor of Engineering degree in Computer Science and Engineering from M S Ramaiah Institute of Technology, Visvesvaraya Technological University in the year 2016. His research areas of interest include Machine Learning and Image Processing.

Dr. Arunalatha J S is a Professor with the Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, India. Her areas of research include Digital Image Processing, Computer Networks, and Artificial Intelligence.

Dr. Venugopal K R is the former Vice Chancellor of Bangalore University, Bengaluru, India. He is an IEEE Fellow and an ACM Distinguished Educator. His professional bodies' memberships are Life Member, ACM; Fellow IEI; Fellow IETE; Life Member, ACS; Life Member, ISTE; and Member, IAE. He has authored and edited 72 books and published over 934+ journals. His research areas are Data Mining, Optical Networks, Ad-hoc Networks, Sensor Networks, and Digital Signal Processing.