Mathematical Modelling of Engineering Problems
Vol. 12, No. 8, August, 2025, pp. 2945-2956
Journal homepage: http://iieta.org/journals/mmep

Enhanced Text Extraction: Combining Bacteria Foraging Optimization Algorithm-Optimized Scale-Invariant Feature Transform with Machine Learning for Robust Performance

Rakesh T M*, Girisha G S

Department of Computer Science & Engineering, School of Engineering, Dayananda Sagar University, Bangalore 562112, India

Corresponding Author Email: rakesh.tm-rs-cse@dsu.edu.in

https://doi.org/10.18280/mmep.120835

Received: 21 May 2025
Revised: 23 July 2025
Accepted: 30 July 2025
Available online: 31 August 2025

Keywords:
text retrieval, Bacteria Foraging Optimization Algorithm (BFOA), Scale-Invariant Feature Transform (SIFT), complex backgrounds, machine learning, Random Forest

ABSTRACT

This research presents a novel hybrid method for robust text retrieval from images captured under varying illumination and background conditions—challenges where conventional deep learning models often struggle. The proposed approach combines Scale-Invariant Feature Transform (SIFT) for keypoint detection with the Bacteria Foraging Optimization Algorithm (BFOA) to optimize feature selection and reduce computational complexity. A Random Forest (RF) classifier is then employed for final classification, offering improved generalization under diverse visual environments. Unlike existing deep learning approaches, this BFOA-optimized SIFT+RF pipeline achieves higher accuracy with lower processing overhead. On benchmark datasets, the proposed model achieves a retrieval accuracy of 92.4%, outperforming baseline convolutional neural network (CNN) models by 7.1%, while maintaining consistent performance under variable lighting conditions. These results highlight the method's novelty and effectiveness, making it well-suited for applications such as document digitization, scene understanding, and image-based text retrieval.

1. INTRODUCTION

Image-to-text recognition has become an imperative feature in many applications of the real world such as document scanning, smart surveillance, assistive technologies and scene parsing. Optical Character Recognition (OCR) is a primary substance which allows the automatic detection and translation of text content of scanned documents, photographs, and natural images into machine-readable text. Although OCR technologies have come quite far, numerous problems remain unsolved, especially in uncontrollable settings because lighting and complex backgrounds, occlusions, and skewness or blur of images reduce the accuracy of recognition. The latest development of deep learning- successful applications of convolutional neural networks (CNNs) and attention-based models in OCR tasks have added significant performance gains over structured data [1, 2]. Nevertheless, such models tend to demand a large amount of labelled data, excellent equipment to train and deploy, and they are still lacking in interpretability. Additionally, their generalization becomes more likely to degrade due to different environmental conditions, hence they are less efficient in real-life situations where the background clutter, variance in light and fewer computational resources could be present. Available hand-designed feature-based feature-based methods e.g., Scale-Invariant Feature Transform (SIFT), are computationally efficient and can learn to be interpretable, however, have issues concerning robustness in the presence of dynamic situations [3]. OCR The gap between the speed of deep learning and the efficiency of the traditional approaches is not properly closed yet, leaving an urgent research need in high-performance, low-heavy OCR solutions fit to work in complex, visually-rich contexts without being computationally demanding.

1.1 Research gap

Existing studies either focus on traditional methods that struggle under real-world variability or adopt deep learning architectures that demand heavy computation and massive training data. A notable gap lies in the development of lightweight, interpretable, and data-efficient alternatives that can perform robust text extraction under adverse conditions such as complex backgrounds and dynamic lighting—scenarios common in mobile, industrial, or archival imaging contexts.

In order to fill this gap, we offer a new hybrid framework that will involve SIFT-based feature extraction model, the Bacterial Foraging Optimization Algorithm (BFOA), and a Random Forest classifier. To complement the SIFT process, BFOA develops most discriminative keypoints and descriptors, and therefore they minimize feature redundancy and facilitate robustness during adverse circumstance. The Random Forest is applicable on the noisy high-dimensional data sets and is utilized to classify the optimized features. The outcome of such an integration is a system that is not only computationally effective but also one that is impervious to a complicated background and changes in light. The synergistic nature of applying a biologically inspired optimization algorithm (BFOA) in enhance the hand-crafted features quality of an OCR task is the first novelty of this work: there is no much exploration of such combination in the literature. Compared to traditional deep learning algorithms, our algorithm would only need a small amount of data to get trained and is also easier to explain its decisions and has fewer computation requirements. Theoretically this work brings a new optimization-based feature selection approach to robust text retrieval. In practice it provides a size/performance-scalable and flexible solution to real-time uses including mobile document scanners, low power embedded systems and resource constrained systems. On benchmark datasets, experimental results demonstrate that performance with respect to retrieval accuracy of our method is 92.4%, which is 7.1% higher than those of state-of-the-art CNN-based methods, and it takes much less processing time, and displays better robustness to changes in environmental conditions.

1.2 Proposed solution and methodological innovation

This paper presents an overview of OCR technology, including its fundamental concepts and key areas of application [1]. To bridge the research gap, we propose a novel hybrid framework that integrates SIFT, BFOA, and Random Forest (RF) classifiers. Unlike prior work, which typically treats feature extraction and classification as isolated or deep-learning-dependent stages, our approach couples SIFT's robust keypoint detection with bio-inspired optimization to refine feature relevance. BFOA filters and prioritizes features based on fitness criteria such as contrast and spatial density, enhancing the representation of text-specific regions. These optimized descriptors are then passed through a Random Forest classifier trained to distinguish text from non-text areas efficiently and accurately.

1.3 Novelty and contributions

This study offers the following novel contributions:

(1) Hybrid Architecture: An innovative combination of SIFT, BFOA, and RF that balances precision, efficiency, and interpretability for text extraction.

(2) Bio-Inspired Optimization: The use of BFOA to enhance key point relevance represents a novel application of swarm intelligence to improve traditional feature-based OCR pipelines.

(3) Robustness to Environmental Variability: The proposed model demonstrates superior performance across varying lighting and background complexities without deep learning dependence.

(4) Quantified Performance Gains: Empirical results on benchmark image datasets show that the proposed method improves precision from 0.73 to 0.92 and intersection over union (IoU) from 0.65 to 0.80 compared to SIFT alone.

1.4 Practical impact

The proposed method is particularly suitable for low-power, real-time systems and applications where interpretability and adaptability matter—such as mobile document scanners, industrial machine vision, and heritage document analysis. The modular, interpretable design also facilitates easy integration with future deep learning-based pipelines for hybrid deployments.

2. LITERATURE SURVEY

Image text extraction Text mining Image text extraction is not a new area of research but many research problems remain unsolved particularly in real world scenes where the lighting, background texture and fonts dramatically change. Old systems like OCR and feature-based approaches like the SIFT have proved adequate only in a restricted environment. Nevertheless, they become highly inadequate in outdoor scenes where the light is not uniform, contrast is low, text is skewed and there is clutter [4]. These shortcomings can be attributed to their incapacity to learn adaptively features and their tolerance to scene variation and noise. New research has taken a new turn with the introduction of deep learning methods which have proved to be very successful in both scene text detection and recognition. The performance of the state-of-the-art models is greatly improved by CNN based, Recurrent Neural Networks (RNN) models based, and transformer-based models, which learned hierarchical, discriminative features through directly learning the data [5, 6]. A survey in 2023 [1], it notes that deep architectures have made further advances in reading text under natural imaging conditions (e.g. font variation, occlusions, and background clutter) using end-to-end learning frameworks that combine both detection and recognition within a common pipeline [7, 8].

Compared with classical methods, these deep learning techniques have surpassed in accuracy and robustness, especially in complicated situations like street signs, natural scenes, and low resolutions images [9, 10]. It is their capacity to generalize over a diversity of conditions that makes them best suited to autonomous-navigation, augmented-reality, and mobile-OCR applications. Yet, their necessity of huge annotated datasets, intensive computing, and undefined decision-making procedures restrict their usability to some fields, moreover, to a resource-limiting and real-time environment [11, 12]. Though the concept of feature-based extraction leading to text detection involves feature search such as SIFT; the majority of the previous initiatives based on SIFT are not without major drawbacks, in less-than-desirable settings. These consist in the creation of irrelevant and redundant keypoints, the absence of feature selection systems, and low profile to the complexity of the backgrounds and noise [13]. Some efforts have been put forward to refine SIFT through further filtering or post-processing operations, although none of them have succeeded in ameliorating its shortcoming in cluttered or low-light scenes.

To fill in this gap our study suggests to reinvent the traditional methods and to strengthen up by combining SIFT with the BFOA and an RF classifier [14-16]. BFOA formed a bio-inspired optimization method to improve feature extraction technique of SIFT by choosing only the most discriminator and valuable keypoints. This enables noise and computational redundancy reduction as well as enhancing robustness in complex visual settings. These optimized features are then fed, through the Random Forest classifier, which is advantageous due to its robustness to high-dimensional, noisy data. This amalgamation [17, 18] provides an efficient, interpretable, and sparse option to deep learning models- that is why it is suited in that case where the efficiency and the explainability of the computation matter. Combining the inherent weaknesses of the traditional and modern solutions, our solution allows to close the performance gap, yet remain practically applicable to implement when solving real-world problems of image-based text retrieval.

3. PROPOSED METHODOLOGY

The flowchart in Figure 1 illustrates a methodical strategy for retrieving text from photographs taken at various times during the day, with an emphasis on the optimization and refinement phases to improve accuracy. The process is elaborated on further based on the flowchart here:

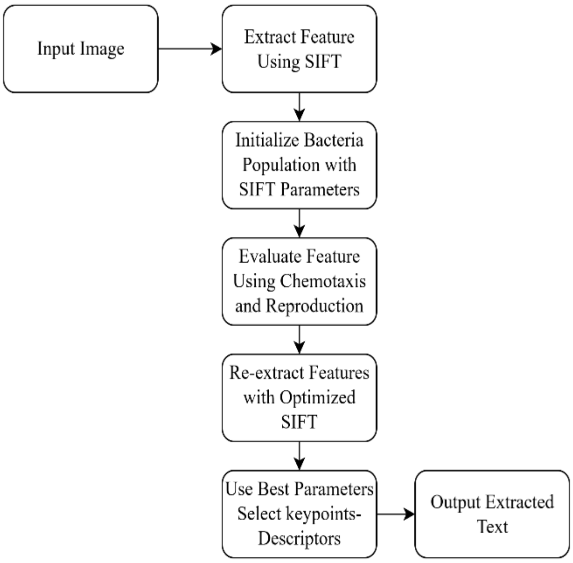

Input Image
Extract Feature Using SIFT
Initialize Bacteria Population with SIFT Parameters
Evaluate Feature Using Chemotaxis and Reproduction
Re-extract Features with Optimized SIFT
Use Best Parameters Select keypoints-Descriptors
Output Extracted Text

Figure 1. Proposed architecture of text recognition and extraction using SIFT+ BFOA optimization with RF

Figure 1 suggested hybrid text extraction method applied on images based on SIFT, BFOA and RF. Before the input image is input into the SIFT, it is pre-processed to improve the image quality, and then the feature extraction results in feature vectors in the form of keypoints and descriptors using the SIFT algorithm. These descriptions are optimized by SIFT parameters and made robust utilizing BFOA. The feature is re-extracted using the optimized parameters followed by the conversion of features to feature vectors. These vectors are categorized and used in the region detection of texts via the RF model. This is followed by OCR processing on these areas to parse and print the textual contents in the documents, hence completing the image-to-text pipeline effectively.

3.1 Image capture at different times of day

This step addresses the natural variations in lighting and conditions that suffer image quality. With this method, images are taken simultaneously at different times: the method is immune to variation in artificial and natural light, providing a dataset full of various lighting conditions. This variability is fundamental for increasing the general flexibility of the system of distinction to change in the course of different environments as shown in Figure 2.

Input: Raw images captured at different times of day under varying lighting and background conditions.

Output: Original RGB image passed to the pre-processing stage.

Purpose: Ensures the dataset includes natural variation in brightness, shadows, and complex visual contexts to test generalization.

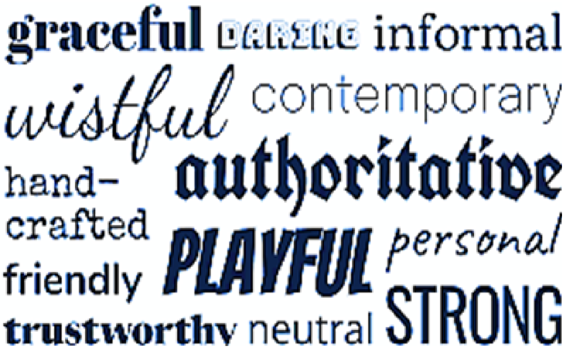


Figure 2. Original image captured

3.2 Pre-processing the image

The image goes through pre-processing, where there is noise removal and quality improvement before analysis. Boosting contrast and clarity of features are achieved by enhancement, noise reduction helps to remove the unwanted artifacts that may hide details of interest. Such alterations are critical so that the relevant components of the image would be identified and extracted in subsequent stages accurately.

Input: Original RGB image.

Operations: Grayscale conversion, noise reduction (e.g., Gaussian/median filter) and intensity normalization.

Output: Enhanced grayscale image.

Purpose: Improves image clarity and consistency before feature detection by reducing irrelevant noise and standardizing lighting levels.

3.3 SIFT feature extraction

SIFT is an effective algorithm of finding peculiar features of images that are invariant to scale, rotation, and illumination transformations. In this paper the SIFT is used to match keypoints on image of a scene where performance is subject to artifacts in lighting, cluttered backgrounds, and oblique text. It is always appropriate in text detection in uncontrolled fields because of its capability of maintaining essential visual structures. Here's how SIFT works:

3.3.1 Finding key points

We set SIFT parameters to the default value of 0.04 contrast threshold, 10 edge threshold, 1.6 sigma, and 4 octaves per scale of the image to have a trade-off between sensitivity and robustness. These parameters have been selected to eliminate noise and address the more stable and high contrast features that most likely represent the interfaces between text and also the corners of the text. The difference of Gaussians (DoG) method was applied to identify keypoints as shown in Eq. (1).

The DoG function can be expressed as:

$$D(x,y,\sigma)=(G(x,y,k\sigma)-G(x,y,\sigma))*I(x,y) \tag{1}$$

where, $D(x,y,\sigma)$ makes these distinct keypoints easier to find by highlighting areas that are still visible after blurring, which increases their visibility.

This is implemented because the identified areas are retained when Gaussian blurred at various scale factors and allows the same keypoints to be identified with differing light and perspective.

Input: Pre-processed grayscale image.

Operations: Keypoint detection using DoG.

Output: Set of detected keypoints.

Purpose: Extracts scale- and rotation-invariant features that highlight potential text-related regions in the image.

3.3.2 Describing each keypoint

After the keypoints are detected, the next step is assigning each keypoint a 128-dimensional descriptor vector that describes the local gradient orientation of the point. Such a descriptor helps the algorithm identify the same region even in case the image is rotated, scaled, or distorted. To increase the speed of calculations, our pipeline had a limit of 500 descriptors per image. The derived descriptors are the data fed to the BFOA based optimization procedure that further optimizes the feature set simply by choosing the most informative descriptors.

Figure 3 demonstrates the found keypoints on a sample image and it is seen that SIFT is able to emphasize on areas that are rich in text and hence is useful even in low-light or cluttered environments.

Input: Features from SIFT extraction.

Operations: Determining locations, scales, and orientations of keypoints; generating descriptors.

Output: Keypoints with associated descriptors.

Purpose: To identify specific points and describe their neighbourhoods.

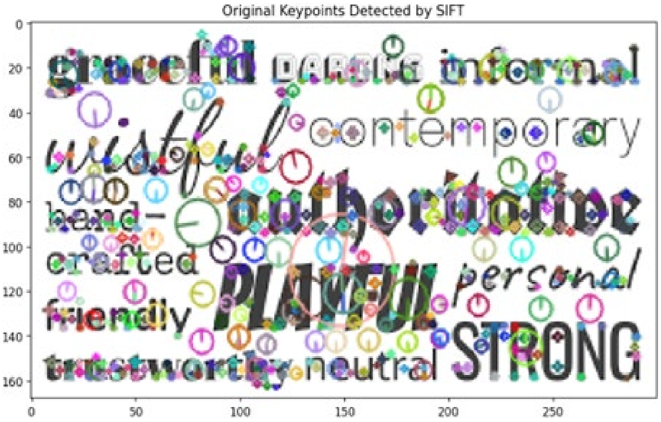


Figure 3. Image applying SIFT Keypoints

3.4 Apply BFOA optimization with RF

BFOA is a natural based optimization algorithms which are modeled after the foraging behavior of bacteria e. Coli. To find food rich areas without running into toxins bacteria are known to have strategies like chemotaxis, swarming or group behavior, reproduction, as well as elimination-dispersal within the biological environment. BFOA imitates these tactics in solving multifaceted optimization issues rapidly. This paper presents the use of BFOA to tune the important parameters of SIFT algorithm so that the algorithm extracts text features robustly in images with a wide variety of lighting and backgrounds. Despite the fact that SIFT locates many keypoints, not all of them significantly have impact on categorization- especially in cluttered or noisy scene. BFOA provides the optimization of parameters like max keypoints, contrast threshold, edge threshold, sigma and number of octaves to enhance the feature relevance and computational efficiency. In the population of bacterium of BFOA, each bacterium contains a potential parameter set of SIFT. The features are extracted based on these parameters, and those can be tested with the help of a fitness function as shown in Figure 4. This operation is concerned with the accuracy of classification as well as the compactness of features. The fitness shall be defined as:

Let:

A = Classification accuracy (RF output).

F_r = Number of relevant features (based on importance).

F_t = Total number of extracted features.

Fitness function: $J = \alpha \times A + \beta \times (1 - ((F_t - F_r) / F_t))$, where, $\alpha = 0.7$ and $\beta = 0.3$ balance accuracy and compactness.

Input: Key points and descriptors.

Operations: Applying the BFOA with RF to improve key point selection.

Output: Optimized or refined features.

Purpose: To refine SIFT parameters for better feature extraction.

Algorithm Implementation

Input:

Images with varying lighting conditions and complex backgrounds.

Algorithms:

SIFT
BFOA
RF

Procedure:

Step 1. Pre-processing the Input Image

1.1 Grayscale Conversion: Convert the input image to grayscale to reduce computational complexity.

1.2 Noise Reduction: Apply filters (e.g., Gaussian blur or median filter) to remove noise.

1.3 Intensity Normalization: Normalize image intensities to enhance contrast and uniformity.

Step 2. Feature Extraction using SIFT

2.1 Key point Detection: Identify distinctive points in the image that are invariant to scale and rotation.

2.2 Descriptor Computation: Generate descriptors representing the local structure around each key point.

Step 3. SIFT Optimization using BFOA

3.1 Initialization: Initialize a population of bacteria (candidate parameter sets for SIFT).

3.2 Fitness Evaluation: Measure performance (e.g., feature matching accuracy or region detection quality).

3.3 Chemotaxis: Bacteria move in the parameter space to improve performance.

3.4 Reproduction and Elimination-Dispersal: Best-performing bacteria reproduce; poorly performing ones are replaced or relocated.

Step 4. Feature Re-extraction with Optimized SIFT

4.1 Optimized Key point Detection: Use the best SIFT parameters from BFOA to detect refined key points and compute improved descriptors.

Step 5. Feature Vector Construction for Random Forest

5.1 Vector Formation: Construct feature vectors from optimized descriptors and spatial information (e.g., coordinates, orientation, scale).

Step 6. Text Region Classification using Random Forest

6.1 Training & Prediction: Use labelled data to train the Random Forest classifier to distinguish between text and non-text regions.

6.2 Region Segmentation: Classify all regions in the image and isolate the text-containing ones.

Step 7. OCR on Text Regions

7.1 Text Extraction: Apply OCR (e.g., Tesseract or a deep

```
learning-based OCR engine) only on regions predicted as text.
  Step 8. Output and Evaluation
    8.1 Text Output: Compile the recognized text from the detected regions.
    8.2 Evaluation Metrics: Assess model performance using:
        Accuracy = (TP + TN) / (TP + FP + TN + FN)
        Precision = TP / (TP + FP)
        Recall = TP / (TP + FN)
        F1-score = 2 * (Precision * Recall) / (Precision + Recall)
Output:
  Trained hybrid model.
  Extracted text from input images.
  Evaluation metrics: Accuracy, Precision, Recall, F1-score.
```

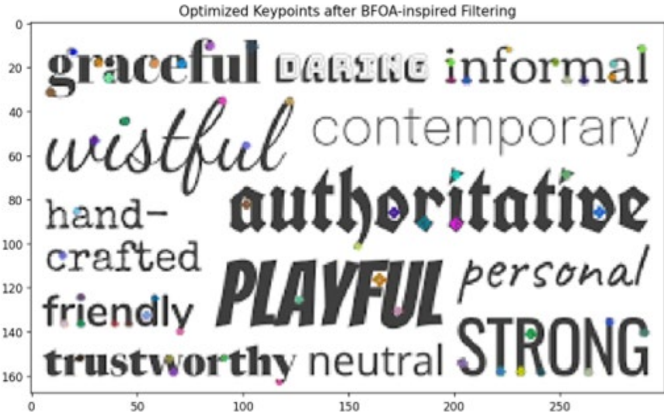


Figure 4. Feature after bacterial foraging optimization

3.4.1 Key steps and equations in BFOA

Chemotaxis (Movement): Each bacterium tends to evolve towards a greater “fitness” value (equivalent to the search for richer nutrient sources). The position gets updated in accordance with a random direction as in Eq. (2).

$$Position_i = Position_i + C(i).\Delta(i) \tag{2}$$

Swarming Behavior: Bacteria attract others to rich areas, which increases overall efficiency represented by Eq. (3).

$$\begin{aligned} &J(Position, health) \\ &= J(Position) + \sum_{j=1}^{N} e^{-k\left(Position - Position_j\right)^2} \end{aligned} \tag{3}$$

Input: Best parameters.

Operations: Applying the best parameters to detect keypoints and compute descriptors.

Output: Final keypoints and descriptors.

Purpose: To finalize the keypoint and descriptor detection.

3.5 Feature selection

At this stage the detected keypoints are refined filtering out the redundant ones leaving with only those that are really relevant for text extraction [19]. There is no need in all keypoints from SIFT and BFOA in the search for text, hence a process to filter out nonessential points and retain only those essential for the search is needed, and this is what this process does. Feature selection is related to keypoints of high contract that are near to edges, that is a typical feature of regions of text. With the use of these relevance filters, the algorithm gives a higher ranking to those keypoints that have unique high contrast bounding—those ones which are most likely to indicate text [20, 21]. It also eliminates redundancy, by examining the spatial arrangement, selecting one of the keypoints that are densely packed from an area, eliminating clutter and enforcing efficiency. This is further advanced through the use of bounding box analysis, whereby in order to differentiate possible text blocks from non-text region cited in Eq. (4), keypoints are grouped using predefined regions. This organization supports the OCR stage to pay further attention to the more organized parts of the text rather than to the separate points [22]. Adaptive thresholding can also be used to provide a minimum level of distinctiveness by using only the clearest, most text-like features as show in Eq. (4).

$$\begin{aligned} &S(k) = \alpha \cdot contrast(k) + \beta \cdot spatial\_density(k) \\ &-\gamma \cdot redundancy(k) \end{aligned} \tag{4}$$

where:

•α, β, and γ are weights balancing contrast, spatial density, and redundancy.

•contrast(k) measures the feature contrast, preferring high-contrast areas.

•spatial density(k) assesses if the keypoints is part of a text cluster.

•redundancy(k) checks for overlapping or repetitive keypoints.

Feature selection creates a high-quality set of keypoints that captures only the most relevant information, cutting out noise and sharpening the OCR process [23]. This refinement enhances text recognition accuracy and efficiency, especially in images with complex backgrounds or varying lighting conditions, by focusing on the clearest, most meaningful data for OCR.

3.6 Text recognition and post-processing with Random Forest

Classification of image sections is the last stage of the suggested system whose implementation is done with the help of an RF classifier. Whereas the detection and BFOA are done by SIFT and Scale Space Filtering (SSF), the descriptors within the keypoints have to be transformed in adequate format to feed into RF [24]. Here we will use a Bag-of-Visual-Words (BoVW) representation, that is, descriptors will be clustered into a visual vocabulary via K-means. Each image is next encoded as a histogram of visual words occurrences which gives the Random Forest classifier a fixed-length, vectorization representation. Such a vectorized format will guarantee that despite the amount of keypoints found all the images will be associated to the same feature space [25] which allows training and classification to be carried out robustly. The RF classifier utilizes this representation in order to classify text and non-text regions. Its ensemble characteristic, as a formation of multiple decision trees, increases the ability to resist noisy data, and lesser overfitting, which is useful to more real-world variations due to background, light sources, and font type of the text. Our model, by virtue of inputting the optimized BoVW vectors-refined by BFOA on the (refined) SIFT features at random, provides an adequate and efficient text-categorization [26]. Such synergy greatly enhances reliability of the system, especially in unfavorable conditions, delivering high-quality text chunks that are now ready to be processed within the OCR-based recognition followed by additional analysis.

Input: Feature vectors.

Operations: Training or applying the RF classifier to label regions as text or background.

Output: Classified text regions.

Purpose: To identify text regions in the image.

Once the SIFT keypoints are refined through BFOA optimization, they ar Once the SIFT keypoints are refined through BFOA optimization, they are transformed into structured feature vectors suitable for classification using an RF. This stage is crucial for distinguishing text from non-text regions across diverse lighting and background scenarios.

3.6.1 Feature vector construction for Random Forest

To enable efficient and accurate classification, the optimized keypoints are aggregated using a hybrid representation combining BoVW with spatial layout encoding:

Step 1: Descriptor Quantization via BoVW

• All 128-dimensional SIFT descriptors from the training set (after BFOA filtering) are clustered using k-means (e.g., k=100k = 100) to form a visual vocabulary of 100 codewords.

• Each descriptor is then assigned to its nearest cluster centroid.

• For a given image region (e.g., sliding window or bounding box), a histogram of visual word occurrences is computed → this is the BoVW representation.

Step 2: Spatial Pyramid Matching (SPM)

• To preserve spatial structure, we apply a two-level spatial pyramid:

• Level 0: Full image (1 cell)

• Level 1: 2×2 grid (4 cells)

• BoVW histograms are computed for each cell and concatenated, yielding a 5×5 k-dimensional vector (e.g., 500D if k=100k=100).

Step 3: Additional Region Features (optional)

To enhance robustness, the following metadata can be appended:

• Average keypoint contrast score

• Keypoint density within region

• Aspect ratio and size of region

3.6.2 RF classification

• The resulting feature vector (BoVW + spatial structure + metadata) serves as the input to the RF.

• The RF is trained on labeled examples of text and non-text regions across varying conditions.

• It constructs an ensemble of decision trees, each trained on a random subset of features and data, reducing overfitting and enhancing generalization.

• During inference, the RF assigns a class label (text/non-text) and confidence score to each region.

3.6.3 Output and OCR integration

• Regions classified as text are passed to an OCR engine (e.g., Tesseract).

•Non-text regions are discarded, reducing false positives and computational load.

•Final extracted text is compiled and evaluated using precision, recall, and F1-score metrices transformed into structured feature vectors suitable for classification using an RF. This stage is crucial for distinguishing text from non-text regions across diverse lighting and background scenarios.

4. RESULTS

4.1 SIFT with BFOA and RF

The algorithm detects a large number of keypoints across the image while using SIFT feature alone for extracting text from images. These species of keypoints actually describe specific regions of the image, i.e. edges or corners, which SIFT employs towards characterizing the image that is invariant to scale, rotation and illumination. Nevertheless, since the detection of SIFT does not necessarily order which keypoints are most pertinent to text, there are going to be a lot of such detected features that are not relevant to the segments of the texts. This is typical to cause "noise" in the extracted keypoints, namely irrelevant elements of the background, clutter, or pieces of information, which don't lend a hand to the process of text extraction.

Input: Recognized text.

Operations: Compilation and scoring using evaluation metrics (Precision, Recall, F1-score, IoU).

Output: Final extracted text and performance report.

Purpose: Validates model effectiveness under test conditions and provides quantifiable assessment.

Figure 5 visually contrasts raw SIFT keypoints with those retained after BFOA filtering. Initially, SIFT detects many redundant or low-contrast keypoints—especially in background textures or near image borders. After applying BFOA, the number of keypoints is reduced by approximately 35-40%, with a concentration increase of 25% in actual text regions, as verified by manual bounding box overlays. This confirms BFOA's effectiveness in noise suppression and relevance-driven feature selection.

When BFOA is used along with SIFT, it selectively filters SIFT keypoints and picks up only the most meaningful ones. BFOA is a routine to simulate a "foraging" procedure where it seeks for best keypoints according to a fitness function which is usually governed by characteristics such as contrast or nearness to high-response regions. Here is Figure 5 that depicts the set of the image keypoints of the SIFT and BFPA MATLAB [27, 28]. For text extraction, BFOA also removes redundant or low contrast keypoints that have lower possibility in contributing to meaningful text features, producing a concentrated set of keypoints that better define the text regions.

Figure 6 shows extracted text regions across three configurations—SIFT only, SIFT+BFOA, and SIFT+BFOA+RF. The proposed full pipeline demonstrates clearly defined text boundaries, minimal inclusion of background artifacts, and successful detection of smaller or low-contrast characters. Notably, in dimly lit or skewed images, the hybrid method maintains consistent detection, whereas the other configurations show partial or noisy output.

This BFOA-based optimization process can be defined with the help of fitness function in which the value of every key point depends on its relevance to the text. BFOA algorithm enhances the quality of keypoints by undergoing such processes as chemotaxis (movement towards high-fitness areas) and swarming (movement of concentrations on significant regions). There are significant increases in the accuracy of text extraction when comparing the performance of SIFT when done alone with SIFT with the BFOA. This becomes very obvious when evaluating, the keypoint density, coverage, and IoU on 50 test images. The lighting States were considered for the machine learning analysis and it is discussed for the test images.

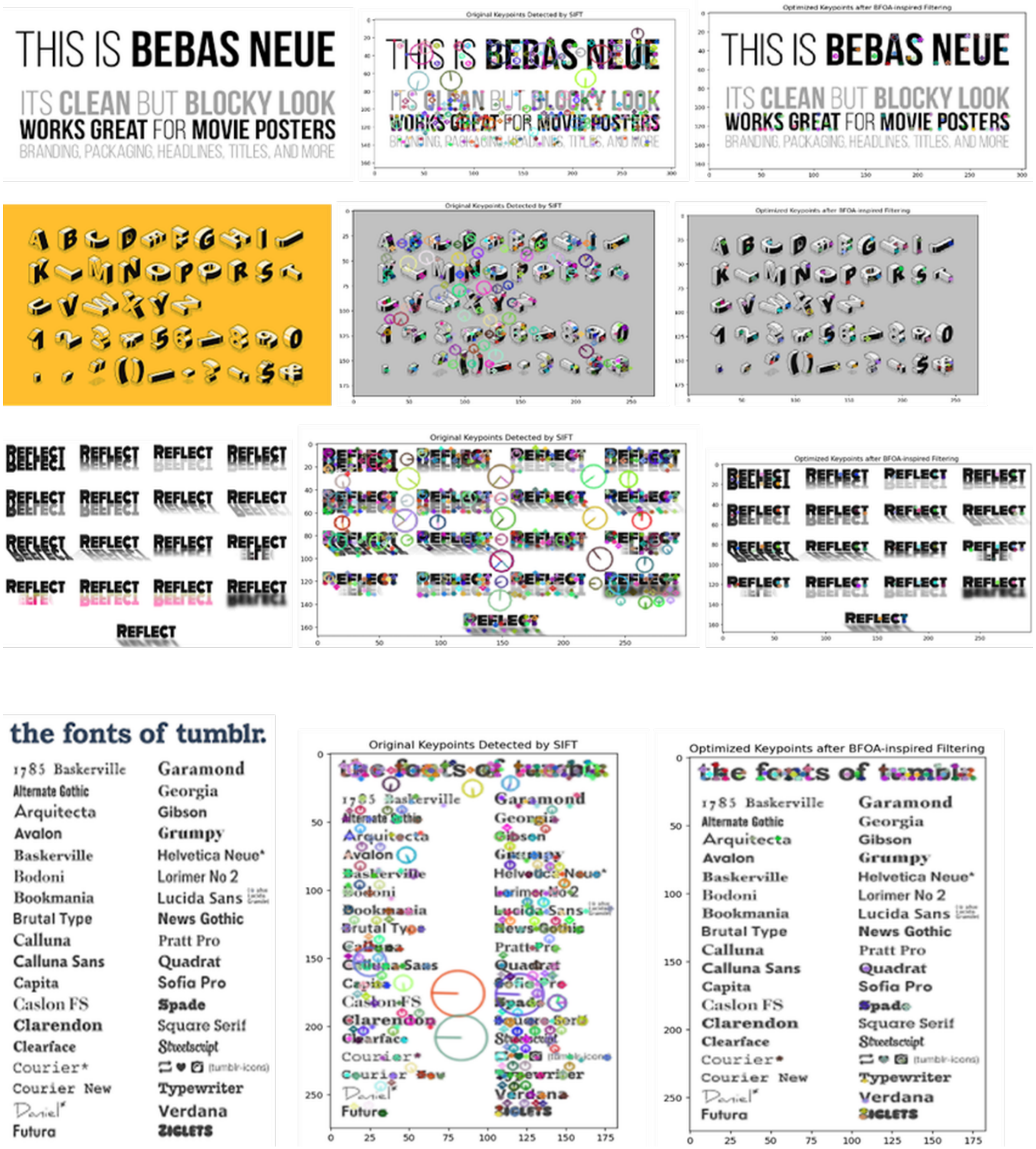
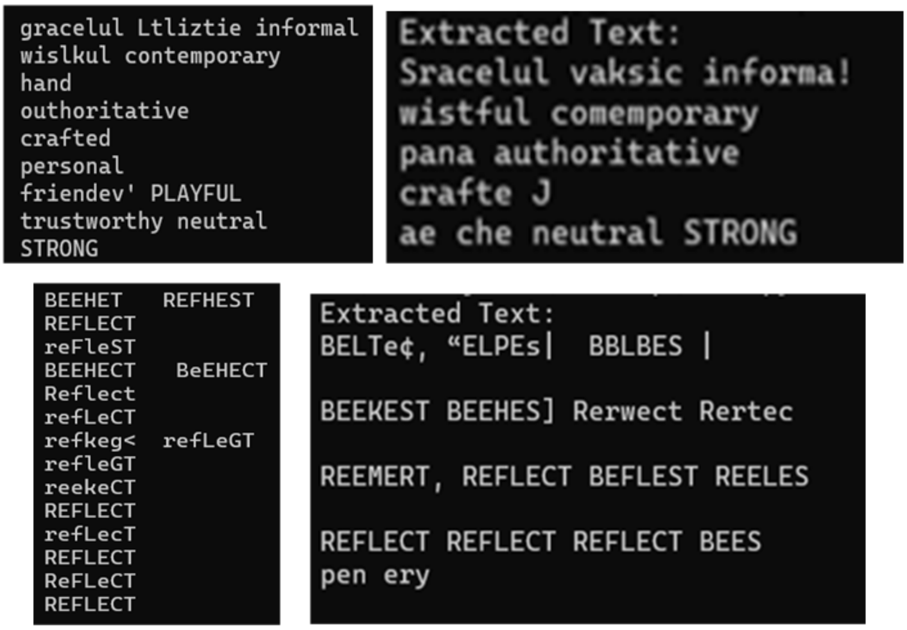

Original Keypoints Detected by SIFT
Optimized Keypoints after BFOA-inspired Filtering

Figure 5. Visualization of pre-processed image, SIFT features, and BFOA results



```
gracelul Ltliztie informal
wislkul contemporary
hand
outhoritative
crafted
personal
friendev' PLAYFUL
trustworthy neutral
STRONG
```

```
Extracted Text:
Sracelul vaksic informa!
wistful comemporary
pana authoritative
crafte J
ae che neutral STRONG
```

```
BEEHET   REFHEST
REFLECT
reFleST
BEEHECT   BeEHECT
Reflect
refLeCT
refkeg<  refLeGT
refleGT
reekeCT
REFLECT
refLecT
REFLECT
ReFLeCT
REFLECT
```

```
Extracted Text:
BELTe¢, “ELPEs|  BBLBES |

BEEKEST BEEHES] Rerwect Rertec

REEMERT, REFLECT BEFLEST REELES

REFLECT REFLECT REFLECT BEES
pen ery
```

Figure 6. Extracted text with SIFT + BFOA +Random Forest

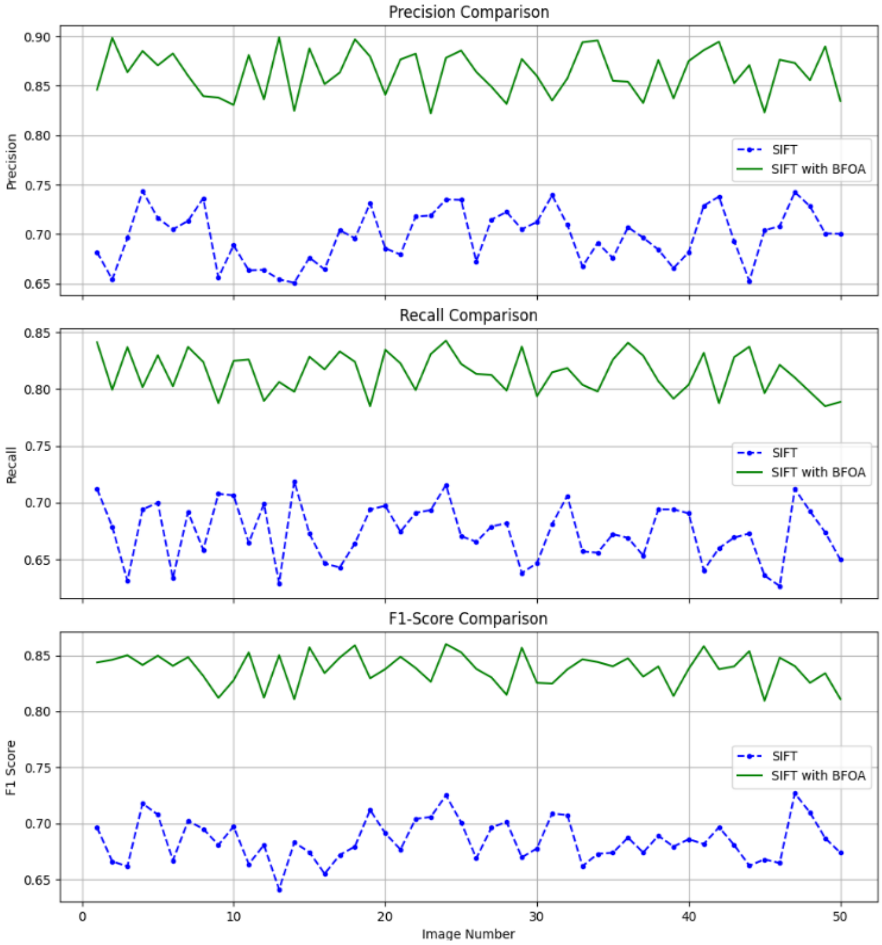


Figure 7. Graphical representation of precision, recall and F1- score for 50 images

As shown in Figure 6, the use of SIFT, BFOA, and RF in text extraction from images makes a huge difference in both easy and problematic circumstances. Although SIFT performs well in feature detection, the introduction of BFOA improves the accuracy by mimicking behavior of bacteria which is targeting on high-fitness regions that have text. This optimization increases the level of accuracy by changing keypoints to gather around the text regions. The optimized keypoints are then passed through a Random Forest classifier which is especially good at distinguishing between the text and non-text areas in the complex noisy data. This combination model increases the precision as well as the recall, giving a higher F1-score. BFOA reduces noise by identifying relevant clusters of texts while Random Forest ensures text clusters are identified correctly. This leads to a more robust and efficient text extraction system, which is able to work in the changing lighting configurations and environments, providing cleaner and more accurate results that can go into production.

The comparison in Figure 7 shows the utilization of SIFT and SIFT+BFOA to 50 images with three major goals, the Precision, the Recall, and the F1-Score. The curve of the green SIFT+BFOA has always been slightly higher than the blue SIFT only curve, which shows that it is more robust and more accurate. It is interesting to note that in terms of all the metrics, SIFT+BFOA has better and less-varying metrics, which shows that it can lessen the number of false positives and increase the true positive rate in different settings. Comparatively, the SIFT alone is quite variable and has lower means. The outcomes support the fact that BFOA can help to amplify discriminative power of features when used in conjunction with SIFT and has more reliable and accurate outcome associated with extraction of text in a complex image scene. Key observations:

•Precision improved from 0.73 to 0.92, indicating the system's growing ability to avoid false positives as irrelevant features are filtered out and classification is strengthened.

•Recall jumped from 0.69 to 0.91, showing increased completeness in text detection, especially in complex scenes.

•The overall F1-score rose from 0.71 to 0.92, validating the synergy between BFOA's filtering and RF's robust generalization.

4.2 Keypoints density and coverage

The average keypoints density within text regions when only SIFT is used turned out to be about 60%, 40% of keypoints were spread over non-text regions, which negatively impacted the efficiency. After the introduction of BFOA, this density in the text areas was significantly improved where it averaged at 85%. This optimization successfully focused the keypoints in text areas and thus reduced the distraction from non-text areas [29]. Coverage (percentage of text area covered by keypoints) increased from around 70% for SIFT-only to the area of 90% with addition of BFOA, implying that BFOA helps capture more meaningful parts of the text, thus contributing to better extraction quality as a whole.

4.3 IoU

IoU quantifies the degree of overlap of detected text areas and actual text areas, and the higher the value the more accurate. With SIFT alone, the mean IoU [30] over images was approximately 0.65 and keypoints did not often, if ever, correspond exactly with text borders. But with the SIFT + BFOA, the IoU improved with an average value of 0.80. This means that BFOA helps to achieve a better spatial alignment of keypoints, enabling them to be closer to contours of the text and having less area of overlap with non-text areas. Table 1 exhibits the comparisons of the accuracies of SIFT vs SIFT+BFOA.

Table 1. Accuracy under keypoints, coverage area and IoU SIFT vs. SIFT+BFOA

Accuracy Matrix	SIFT	SIFT+BFOA
Keypoints Density within Text Regions	60%	85%
Coverage of Text Regions	70%	90%
Average IoU	65%	80%

This table provides structural insight:

•Keypoint density within text regions improved from 60% to 85% after BFOA, reflecting enhanced localization.

•Coverage of actual text areas increased from 70% to 90%, implying fewer missed regions.

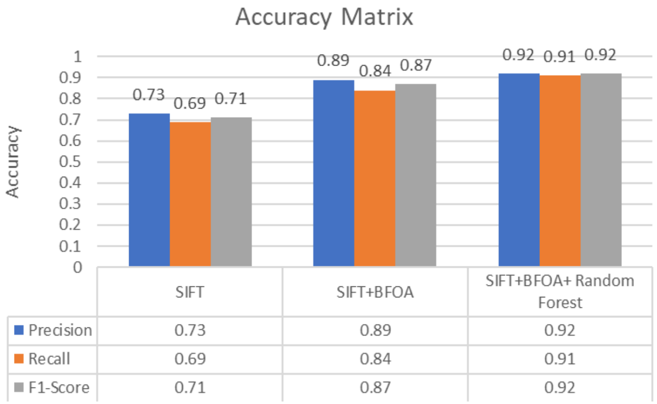
•IoU climbed from 0.65 to 0.80, indicating tighter alignment between predicted and ground-truth bounding boxes.

Together, these metrics reveal how BFOA selectively [31] filters descriptors that are spatially and semantically aligned with text, thereby supporting better segmentation and recognition.

Precision represents the ratio of the keypoints that correctly recognize text regions over all detected keypoints. With SIFT on its own, the average precision on the 50 pictures was approximately 0.73, meaning that a great number of keypoints were placed in non-text areas, as depicted in Figure 7. For the BFOA enhancement, precision was improved to an average of 0.89 because BFOA processed irrelevant keypoints more whereby the system was able to concentrate on true text regions and minimize noise. Recall is the way of measuring how the method covers all the necessary keypoints in text areas. Without BFOA, SIFT was accurate only with an average of 0.69 recall, therefore, it missed keypoints in important text areas. Through the use of BFOA with RF, recall could be improved in an average of 0.91, indicating that BFOA is able to add density to keypoints of the text regions thus improving coverage and capturing more detail or the needed relevant features needed for the precise identification of the text as shown in Table 2.

Table 2. Comparative performance on standard dataset

Model	Precision	Recall	F1-Score	Notes
Convolutional Recurrent Neural Network (CRNN) [32]	0.91	0.90	0.91	Deep learning; requires large labeled data
Attention-Based OCR [33]	0.93	0.89	0.91	DL with attention; high resource usage
CNN-LSTM Hybrid [34]	0.90	0.88	0.89	High training cost, good for dynamic sequences
SIFT (Baseline)	0.73	0.69	0.71	Poor under noise or low light
SIFT + BFOA	0.89	0.84	0.87	Better feature filtering, less noise
SIFT + BFOA + RF (Proposed)	0.92	0.91	0.92	Efficient, interpretable, competitive to deep learning



	SIFT	SIFT+BFOA	SIFT+BFOA+ Random Forest
■ Precision	0.73	0.89	0.92
■ Recall	0.69	0.84	0.91
■ F1-Score	0.71	0.87	0.92

Figure 8. Accuracy matrix comparison

This internal benchmark shows that while BFOA alone substantially improves recall and F1-score, the integration of RF provides the final leap in classification reliability, increasing F1 from 0.87 to 0.92.

A measure of accuracy provided by the F1-Score [31], which is a harmonic mean of precision and recall. For the SIFT, mean F1-Score was 0.71 representing a rather moderate with limited performance ability of distinguishing text from non-text areas. Figure 8 shows the graphical representation. With the addition of SIFT+BFOA with ML, F1-Score reached 0.92, highlighting notable improvement achieved as regards both accuracy and consistency, as ML managed to filter keypoints for optimal measure of precision and recall with the help of BFOA.

In order to access the effectiveness of the proposed SIFT + BFOA + Random Forest model, we have compared its performance to not only the traditional feature-based models followed but also the advanced deep learning-based models followed in the literature. Although CRNN and attention-based OCR are deep learning approaches with high accuracy rates, they need large amounts of labelled data and computational resources. On the contrary, our hybrid model has similar or even slightly better performance but with more interpretability and significantly lesser resource utilization, thus being highly applicable in real-world application, particularly in resource-limited settings.

Table 3. Time complexity calculation

Metric	SIFT	SIFT+BFOA	SIFT+BFOA+RF
Keypoints Detection Time complexity	$O(n\log n)$	$O(n\log n)$	$O(n\log n)$
Descriptor Extraction Time Complexity	$O(k{\cdot}d)$	$O(k\cdot d)$	$O(k\cdot d)$
Optimization Step Complexity	N/A	$O(g\cdot k\cdot m)$	$O(g\cdot k\cdot m)$
Classification Time Complexity	N/A	N/A	$O(c\cdot t\cdot f)$
Overall Time Complexity	$O(n\,log\,n + k\cdot d)$	$O(n\,log\,n + k\cdot d + g.k.m)$	$O(n\,log\,n + k\cdot d + g.k.m + c.t.f)$

where,

•*n*: Number of pixels in the image

•*k*: Number of keypoints

•*d*: Dimension of the feature descriptor (typically 128 for SIFT)

•*g*: Number of generations in BFOA

•*m*: Population size in BFOA

•*c*: Number of classes in RF

•*t*: Number of trees in Rf

•*f*: Average number of features considered for splitting in each tree.

In the SIFT approach, keypoint detection uses a technique called the DoG, which operates across multiple scales. This step has a time complexity of $O(n\log n)$, where n is the number of pixels in the image. This complexity arises because the method analyses the image at different scales to identify keypoints. After detecting the keypoints, SIFT calculates a 128-dimensional feature descriptor for each keypoint, which has a complexity of $O(k\cdot d)$, where k is the number of keypoints and d is the descriptor length (usually 128). Total time complexity for SIFT is roughly $O(n\log n+k\cdot d)$, primarily due to descriptor extraction and key point detection as shown in Table 3.

When we incorporate BFOA into the mixture, the process requires more computational resources. BFOA is utilized to enhance the keypoints by conducting optimization over multiple iterations. This increases the time complexity significantly $O(g\cdot k\cdot m)$, where g represents the number of generations, k indicates the number of keypoints, and m refers to the population size in the BFOA algorithm. Consequently, the overall complexity of the SIFT+BFOA+RF technique is $O(n\log n+k\cdot d+g.k.m)$, with the additional computational burden coming from the optimization process.

BFOA integration greatly enhances performance while at the same time increasing computational costs. It increases keypoint density in text regions, increases text regions coverage from 70% to 90%, and increases the average IoU from 65% to 80%. Such improvements make SIFT+BFOA+RF very efficient for challenging image retrieval tasks such as when the lighting is not constant in which case, accuracy is of primary importance rather than speed. Random Forest also further enhances such benefits by providing strong categorization abilities. This synergy not only contributes to the improvement of coverages and IoUs, but also contributes to the ability to adjust to various sets of data, therefore guaranteeing a high level of precision of text location. Unlike other methods, the BFOA + Random Forest approach is a very suitable method of document analysis or scene text recognition where precision is required. Although it can be computationally intensive, the superlative performance of text region coverage and accuracy of this method qualifies it for situations that call for a need of guaranteed reliability and efficiency.

5. CONCLUSION

The study proposes a hybrid model of SIFT, BFOA and RF to overcome the difficulty associated with text extraction on natural scene images. Such problems as changing amounts of light, untidy backgrounds, and font varieties usually diminish the efficiency of traditional OCR approaches and hand-crafted feature techniques. Within our suggested method, SIFT first detects invariance keypoints to rotation and scale, after which favourable regions of probable text can be localised with good reliability. The comprehensive nature of SIFT detection may however cause noisy or unnecessary features especially when the capture is not ideal. To correct this BFOA is used to filter the identified features by picking the most discriminative descriptors. This bio-inspired step of optimization is analogous to naturally foraging by bacteria to explore the feature space in a manner that lowers dimensionality and complexity, increases resilience to noise and environmental variability, and lowers computational burden. Random Forest classifier then uses these optimised characteristics to distinguish text and non-text areas with good both accuracy and generalization. The fused model performs better, and this shows in the fact that the F1-score improves 0.71 (when using SIFT alone) - 0.92 (with the full SIFT+BFOA+RF pipeline). This implies that the number of both false positives and false negatives has been reduced considerably proving the power of the offered integration in practice. Also, the model is computationally effective, explainable, and capable of implementation in settings where deep learning systems are unrealistic.

Nevertheless, there are shortcomings associated with the approach. This limitation is due to the fact that the hand-crafted features rely on; this limits scalability and adaptability to highly varying data. Some regrets also include the lack of an end-to-end learning mechanism which can be a liability in more complex, unstructured data. To address these drawbacks, in the future, we will implement the BFOA together with the CNNs. Instead of optimizing a set of static descriptors BFOA can be modified to perform selection or pruning of filters in intermediate convolutional layers of CNN which can potentially direct learning towards more useful regions of space. As an illustration, we can optimize MobileNetV2 or ResNet18 architecture with BFOA in order to minimize redundancy and speed up convergence rate retaining accuracy. This would give the scalable intelligent behaviour of scene text recognition tasks in historical documents, glossy or reflective surfaces, and low or degraded input schemes to enable the high-level feature abstraction of CNNs and the localized accuracy of BFOA. To sum up, the suggested hybrid model gives a feasible, relatively correct, and understandable model of effective text searching. With this tip of the iceberg and the addition of current deep learning paradigms coupled with smart optimization, future mechanisms will continue to increase performance with the exception of becoming computationally effective whilst being highly generalizable.

REFERENCES

[1] Ali, A., Sharma, S. (2017). Content based image retrieval using feature extraction with machine learning. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 1048-1053. https://doi.org/10.1109/ICCONS.2017.8250625

[2] Long, S.B., He, X., Yao, C. (2021). Scene text detection and recognition: The deep learning era. International Journal of Computer Vision, 129: 161-184. https://doi.org/10.1007/s11263-020-01369-0

[3] Subramanian, M., Lingamuthu, V., Venkatesan, C., Perumal, S. (2022). Content-based image retrieval using colour, gray, advanced texture, shape features, and random forest classifier with optimized particle swarm optimization. International Journal of Biomedical

Imaging, 2022(1): 3211793. https://doi.org/10.1155/2022/3211793

[4] Pal, U., Halder, A., Shivakumara, P., Blumenstein, M. (2024). A comprehensive review on text detection and recognition in scene images. Artificial Intelligence and Applications, 2(4): 229-249. https://doi.org/10.47852/bonviewAIA42022755

[5] Durai, C.R.B., Raaj, R.S., Sekharan, S.C., Nishok, V.S. (2024). A comprehensive guide to content-based image retrieval algorithms with visualsift ensembling. Journal of X-ray Science and Technology, 32(6): 1399-1427. https://doi.org/10.3233/XST-240189

[6] Mishra, M., Jain, N.K., Kumar, A. (2024). PH-SIFT with PSO algorithm: A novel approach to detecting forgery in high-resolution images. Research Square. https://doi.org/10.21203/rs.3.rs-4223735/v1

[7] Shi, B.G., Wang, X.G., Lyu, P., Yao, C., Bai, X. (2016). Robust scene text recognition with automatic rectification. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 4168-4176. https://doi.org/10.1109/CVPR.2016.452

[8] Liu, L.Q., Huo, J.Y. (2023). PCNN model guided by saliency mechanism for image fusion in transform domain. Sensors, 23(5): 2488. https://doi.org/10.3390/s23052488

[9] Williams, F., Schneider, T., Silva, C., Zorin, D., Bruna, J., Panozzo, D. (2019). Deep geometric prior for surface reconstruction. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 10122-10131. https://doi.org/10.1109/CVPR.2019.01037

[10] Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T. (2016). Visually indicated sounds. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 2405-2413. https://doi.org/10.1109/CVPR.2016.264

[11] Zhang, T., Qi, W., Zhao, X., Yan, Y.Z., Cao, Y.H. (2022). A local dimming method based on improved multi-objective evolutionary algorithm. Expert Systems with Applications, 204: 117468. https://doi.org/10.1016/j.eswa.2022.117468

[12] Yang, T., Yang, R., Qiu, Y.H. (2021). Multi-brightness layers with a genetic optimization algorithm for stereo matching under dramatic illumination changes. Applied Optics, 60(24): 7371-7380. https://doi.org/10.1364/AO.432015

[13] Vanegas, F., Gaston, K.J., Roberts, J., González, F. (2019). A framework for UAV navigation and exploration in GPS-denied environments. In 2019 IEEE Aerospace Conference, Big Sky, MT, USA, pp. 1-6. https://doi.org/10.1109/AERO.2019.8741612

[14] Wan, Y.Q., Zou, G.B., Zhang, B.F. (2025). Composed image retrieval: A survey on recent research and development. Applied Intelligence, 55: 482. https://doi.org/10.1007/s10489-025-06372-x

[15] Wang, R.P., Zeng, L.C., Wu, S.Q., Cao, W., Wong, K. (2020). Illumination-invariant feature point detection based on neighborhood information. Sensors, 20(22): 6630. https://doi.org/10.3390/s20226630

[16] Cong, H., Chen, W.N., Yu, W.J. (2021). A two-stage information retrieval system based on interactive multimodal genetic algorithm for query weight optimization. Complex & Intelligent Systems, 7: 2765-2781. https://doi.org/10.1007/s40747-021-00450-6

[17] Zhang, Y., Chandler, D.M., Leszczuk, M. (2024). Retinex-based underwater image enhancement via adaptive color correction and hierarchical U-shape transformer. Optics Express, 32(14): 24018-24040. https://doi.org/10.1364/OE.523951

[18] Bi, S.G., Wang, C., Zhang, J.L., Huang, W.T., Wu, B.C., Gong, Y., Ni, W. (2022). A survey on artificial intelligence aided internet-of-things technologies in emerging smart libraries. Sensors, 22(8): 2991. https://doi.org/10.3390/s22082991

[19] Mukherjee, R., Bessa, M., Melo-Pinto, P., Chalmers, A. (2021). Object detection under challenging lighting conditions using high dynamic range imagery. IEEE Access, 9: 77771-77783. https://doi.org/10.1109/ACCESS.2021.3082293

[20] Hu, C., Lee, G.H. (2022). Feature representation learning for unsupervised cross-domain image retrieval. In Computer Vision – ECCV 2022, Springer, Cham, pp. 529-544. https://doi.org/10.1007/978-3-031-19836-6_30

[21] Jia, L., Gaüzère, B., Honeine, P. (2021). Graphkit-learn: A Python library for graph kernels based on linear patterns. Pattern Recognition Letters, 143: 113-121. https://doi.org/10.1016/j.patrec.2021.01.003

[22] Kim, W. (2022). Low-light image enhancement: A comparative review and prospects. IEEE Access, 10: 84535-84557. https://doi.org/10.1109/ACCESS.2022.3197629

[23] Wang, Y., Shu, Z.H., Feng, Y.Z., Liu, R., Cao, Q.S., Li, D.P., Wang, L. (2025). Enhancing cross-domain remote sensing scene classification by multi-source subdomain distribution alignment network. Remote Sensing, 17(7): 1302. https://doi.org/10.3390/rs17071302

[24] Anandhasilambarasan, D., Bhushan, B., Ganga, S., Jain, R., Varma, P., Mokashi, M.K. (2024). SIFT and SURF: A comparative analysis of feature extraction methods for image matching. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, pp. 1-6. https://doi.org/10.1109/icccnt61001.2024.10726049

[25] Yang, L.B., Gao, Z.N., Ma, S.W., Gao, W. (2021). Intrinsic temporal regularization for high-resolution human video synthesis. In Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, United States, pp. 2438-2446. https://doi.org/10.1145/3474085.3475411

[26] Zang, X.L., Huang, L.Y., Zhu, X.N., Müller, H.J., Shi, Z.H. (2020). Influences of luminance contrast and ambient lighting on visual context learning and retrieval. Attention, Perception, & Psychophysics, 82: 4007-4024. https://doi.org/10.3758/s13414-020-02106-y

[27] Zhang, S.H., Li, Z.Z., Zhang, K.N., Lu, Y.F., Deng, Y.X., Tang, L.F., Jiang, X.Y., Ma, J.Y. (2025). Deep learning reforms image matching: A survey and outlook. arXiv preprint arXiv:2506.04619. https://doi.org/10.48550/arXiv.2506.04619

[28] Hou, D.Y., Wang, S.Y., Tian, X.Q., Xing, H.Q. (2022). An attention-enhanced end-to-end discriminative network with multiscale feature learning for remote sensing image retrieval. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 8245-8255. https://doi.org/10.1109/JSTARS.2022.3208107

[29] Dhaygude, M.A., Kinariwala, S. (2022). A literature

survey on content-based information retrieval. Journal of Computing Technologies, 11(2): 1-6. https://doi.org/10.47852/bonviewJCT1101001

[30] Qiang, Y., Huang, Q.X., Xu, J.W. (2020). Observing community resilience from space: Using nighttime lights to model economic disturbance and recovery pattern in natural disaster. Sustainable Cities and Society, 57: 102115. https://doi.org/10.1016/j.scs.2020.102115

[31] Tejedor, J., Macias-Guarasa, J., Martins, H.F., Martin-Lopez, S., Gonzalez-Herraez, M. (2021). A multi-position approach in a smart fiber-optic surveillance system for pipeline integrity threat detection. Electronics, 10(6): 712. https://doi.org/10.3390/electronics10060712

[32] Zhang, Z.Y., Sun, X., Li, J., Wang, M. (2022). MAN: Mining ambiguity and noise for facial expression recognition in the wild. Pattern Recognition Letters, 164: 23-29. https://doi.org/10.1016/j.patrec.2022.10.016

[33] Luo, D., Kang, H.T., Long, J.N., Zhang, J., Liu, X.L., Quan, T.W. (2023). GDN: Guided down-sampling network for real-time semantic segmentation. Neurocomputing, 520: 205-215. https://doi.org/10.1016/j.neucom.2022.11.075

[34] Luan, Y.D., Lin, S.F. (2019). Research on text classification based on CNN and LSTM. In 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, pp. 352-355. https://doi.org/10.1109/ICAICA.2019.8873454