

Dimensionality reduction for off-line object recognition and detection using supervised learning

Sari Awwad¹, Ahmad Al-Rababa², Salah Taamneh¹, Subhieh M. El-Salhi³, Ala Mughaid⁴

¹Department of Computer Science and Applications, Faculty of Prince Al-Hussein bin Abdullah II of Information Technology, The Hashemite University, Zarqa, Jordan

²Department of Computer Science, Faculty of Information Technology, World Islamic Sciences and Education University (WISE), Amman, Jordan

³Department of Computer Information Systems, Faculty of Prince Al-Hussein bin Abdullah II of Information Technology, The Hashemite University, Zarqa, Jordan

⁴Department of Information Technology, Faculty of Prince Al-Hussein bin Abdullah II of Information Technology, The Hashemite University, Zarqa, Jordan

Article Info

Article history:

Received Mar 14, 2024

Revised Jun 6, 2024

Accepted Jun 24, 2024

Keywords:

Dimensionality reduction PCA
Fisher encoding
Local SIFT features
Object recognition and detection
Supervised learning

ABSTRACT

Object recognition and detection is an area of study, within intelligence and computer vision. It finds applications in fields such as surveillance, detailed activity analysis, robotics and object tracking. The primary focus of research papers in this domain revolves around enhancing the precision of object identification and detection regardless of whether the objects are located indoors or outdoors. To address this challenge, a new approach involving the utilization of SIFT features for information extraction has been proposed. Our approach encompasses two components; the implementation of dimensionality reduction through principal component analysis (PCA) to eliminate redundancies; secondly the incorporation of feature vector encoding using fisher encoding techniques. The RGB-D dataset employed contains 300 objects across scenarios with emphasis on colored aspects rather than depth. The SIFT features are categorized using a support vector machine (SVM) into 7 classes. When compared to using SIFT features integrating them with encoding methods notably enhances recall, precision and F1-score by than 30% through fisher encoding and PCA techniques. The study concludes with an evaluation based on n-cross validation methodology along, with detailed experimental results.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Sari Awwad

Department of Computer Science and Applications

Faculty of Prince Al-Hussein bin Abdullah II of Information Technology, The Hashemite University

Zarqa, Jordan

Email: sari@hu.edu.jo

1. INTRODUCTION

Recognizing and identifying objects are components, in a variety of artificial intelligence applications spanning different fields. Here are some important reasons why they hold significance. Applications in visual technology; object recognition and identification are elements in visual technology tasks like categorizing images, tracking objects, understanding scenes and retrieving images. These functions are crucial in sectors such as self driving vehicles, surveillance setups, medical imaging, robotics and augmented reality [1]. Self driving vehicles; in cars and similar self driving vehicles object recognition and identification allow the vehicle

to recognize and comprehend its environment, including vehicles, pedestrians, traffic signals and obstacles. Precise detection and recognition are vital for navigation and decision making [2].

Researchers often grapple with the challenge of enhancing the precision of object recognition and detection. Some focus, on refining these processes by leveraging classifiers known for their efficacy in areas like computer vision research, such, as activity tracking and object recognition. In contrast other methodologies achieve this goal by manipulating images or adjusting the order of their techniques [3], [4]. When others concentrate on handling scenes figuring out the window sizes, for recognized objects various learning methods time needed for computation ways to improve performance and picking features [5]. Hence, the primary challenge lies in enhancing and augmenting the precision of object recognition and detection through the utilization of dimensionality reduction techniques. This paper discusses the importance of dimensionality reduction and encoding methods, in object detection and recognition which have garnered interest, within the computer vision field [6], [7]. The goal of dimensionality reduction is to create a version of the data that retains its characteristics making it easier to process and enhancing accuracy, in recognizing patterns [8]. When it comes to recognizing objects using RGB-D technology supervised learning driven techniques, for reducing dimensions have demonstrated encouraging results.

This paper presents a concept that involves two contributions:

- Enhancing the precision of object recognition and detection by integrating dimensionality reduction, with encoding an approach compared to studies that predominantly focused on dimensionality reduction alone.
- The proposed method combines detection and recognition which diverges from existing research where the emphasis has typically been on either object detection or recognition but not, in conjunction.

This research aims to answer the following questions:

- What is the extend of dimensionality reduction role in the outcomes of object recognition and detection?
- Does combining dimensionality reduction with encoding increase the accuracy of object recognition and detection?

Moreover, changing the used technique while using object recognition or detection had a positive impact on the studied results. Therefore, the proposed method has been compared with techniques that reduce dimensions and evaluated its effects, on accuracy, speed and memory usage for recognition and detection tasks. Furthermore Authors looked into how easy it's to understand the simplified representation and how well it handles changes in viewpoint lighting and objects blocking the view. The following points summarize our approach, which will be explained in details in section 3.

- Feature extraction: initially, local features are extracted from the images using SIFT features. These features represent distinctive regions or key points in the image.
- Fisher vector encoding: the fisher vector encoding is the key step in this technique. It involves computing the gradient of the log-likelihood with respect to the parameters of a gaussian mixture model (GMM) that models the distribution of visual features assigned to each object. This gradient encodes the statistical information about the distribution of features within each object.
- Dimensionality reduction: the fisher vectors tend to be high-dimensional due to the large number of parameters in the GMM. To reduce the dimensionality and remove redundancy, principal component analysis (PCA) has been applied.
- Supervised learning: the proposed method utilized M-SVM² for class classification to identify the object, among the seven categories discussed in this paper. Additionally, a binary SVM classifier has been employed to detect objects within a scene.

The structure of this paper consists of the following sections:

- 1 INTRODUCTION; this section gives an overview of object recognition and detection outlining the research objectives.
- 2 LITERATURE REVIEW; background information, on methods, techniques, classifiers, and features that are related to research area.
- 3 METHOD; explain the proposed approach in details step by step.
- 4 RESULTS and DISCUSSION; experimental results are detailed here including discussions on the experiments conducted and their results.
- 5 CONCLUSION.

2. LITERATURE REVIEW

This section reviewed studies that concentrate on classifiers, characteristics and kinds of datasets utilized for object recognition and detection. Various methods have been introduced across a range of research fields. Studies concerning detection tackle challenges based on the applications linked to the research [9], [10]. This section has been condensed the research into three parts starting with an overview of the existing studies, on dimensionality reduction methods, for recognizing objects offline.

The method presented by [11] introduces an enhancement to PCA implemented with neural networks. Their approach adds an algorithm that can precisely determine the optimal number of meaningful principal components for data streams. In other words, they expanded neural network PCA with an algorithm that can adjust the dimensionality in large increments at each time step. This approach makes use of neural network PCA's built-in features. Using a variety of datasets, they assessed this method for adaptive dimensionality modification in neural network PCA. The evaluation parameters included convergence speed and accuracy. Various halting guidelines were investigated. The algorithm's ability to forecast n minus m eigenvalues with accuracy proved its strength. This enabled the quality of the results to be further improved. In all but one dataset and parameter combinations, the dimensionality was effectively estimated using the cumulative percentage of total variance criterion. All things considered, the suggested method for adaptive dimensionality adjustment and eigenvalue estimation worked well for online PCA and correctly estimated the ultimate dimensionality well in advance of all data points becoming visible.

Nguyen *et al.* [12] put forward a model to make the process of manually scoring large volumes of handwritten mathematical expressions (HMEs) more efficient and consistent. They suggested that grouping similar HME responses together through clustering could be beneficial. However, there are challenges when it comes to grouping expressions HMEs that involve identifying the locations and symbols, within an HME image as well as determining the distance between two HME images. To overcome this issue a method was introduced using convolutional neural networks (CNNs) to capture symbol representations within an HME. By leveraging a scale CNN approach it becomes possible to pinpoint and categorize symbols of varying sizes effectively. The focus on symbols combined with training improves the accuracy of classification and localization. Additionally a multi level spatial distance metric was proposed to aid in grouping two HME representations. Experimental tests were conducted using the CROHME datasets from 2016 and 2019 yielding results, with purity scores reaching 0.99 and 0.96 respectively.

Artificial intelligence and deep learning methods are now being used for critical e-commerce applications. However, human computing and computer aided design cannot understand different online and offline products. So it's hard for customers to find products like groceries, fashion and health items. Overcoming this limitation of human perception is a major challenge. Mohammad *et al.* [13] research, proposed an advanced fully convolutional neural network (FCNN) deep learning model with global threshold is proposed. Digital and online images are selected, preprocessed and classified. Segmentation is first applied, then classification is done with FCNN. Finally, high performance of 98.7% accuracy, 98.7% sensitivity, and 99.23% throughput is achieved, outperforming current technology.

Tarawneh *et al.* [14] utilized content based image retrieval (CBIR) to search for images based on their visual content. CBIR has been a subject of research, for years with methods developed to extract valuable image features that depict image content. Recently deep learning has gained attention in the field of computer vision in applications related to CBIR. Their study offers an analysis of features used in CBIR systems, including high level deep learning features and low level features like SIFT, SURF, HOG, LBP, and LTP. They assess the effectiveness of CBIR systems with deep learning features compared to low level features by employing different dictionaries and coefficient learning techniques. Furthermore they compare these features with commonly used ones such as Gabor features and color histograms. They also delve into how deep features can differentiate between scenarios using diverse similarity metrics and validation techniques. By utilizing PCA, DWT, and DCT methods they explore how reducing the dimensionality of features impacts CBIR performance. Interestingly when using VGG 16 FC7 features from the Corel 1000 and Coil 20 datasets, with 10 D and 20 D K SVD respectively their findings demonstrate high mean average precision (95% and 93%).

Saleem *et al.* [15] research paper has examined studies, on recognizing activities. Along with introducing techniques for this purpose these surveys have pointed out the trends in human activity recognition across contexts and applications. Human activity recognition plays a role in aspects of modern life that rely on technology, such as human computer interaction, security monitoring, healthcare monitoring, robotics, information retrieval and surveillance. Due to the pace of advancements trends in human activity recognition

evolve quickly demanding a current and comprehensive perspective. This review outlines a classification of approaches to human activity recognition that includes online/ methods, multimodal or single modal techniques well as handcrafted and learning based approaches. The surveys objective is to showcase the spectrum of human activity recognition (HAR) within application domains, types of activities, task complexities, benchmark datasets used and methodologies applied. It provides an analysis of cutting edge HAR methods along with discussions on employed datasets. The studies selected are categorized based on the proposed classification system while examining their characteristics like activity intricacy, dataset sizes and recognition accuracies. This comparative assessment of HAR strategies also brings attention to the challenges faced in this field. Suggests avenues, for future research endeavors.

The second part is related to supervised learning algorithms commonly used in object recognition. Sohoni *et al.* [16] proposed a simple but powerful semi-supervised learning method for detecting visual objects called STAC, which uses data augmentation. STAC leverages very confident pseudo labels for localized objects in an unlabeled image to update the model by enforcing consistency through robust augmentations. The researchers suggested experimental protocols for evaluating semi-supervised object detection. STAC improved AP0.5 from 76.30 to 79.08.

Numerous frameworks, including CNNs, deep belief networks (DBNs), and autoencoders (AEs), have been developed as a result of the quick advancement of deep learning techniques. Given this, the last ten years have seen a significant advancement in the recognition of aquatic objects. Wang *et al.* [17] review paper focuses exclusively on comprehensively reviewing deep learning-based object recognition techniques for both surface and underwater targets. First, common architectures and important concepts are compiled into a single framework to facilitate a comprehensive review. As a result, a large collection of widely used and benchmark datasets for the identification of marine objects is assembled, and a thorough analysis and comparison of deep learning techniques is conducted. Furthermore, there is a lot of discussion on experimental results and potential directions for marine object recognition. Lastly, findings regarding the state-of-the-art in deep learning-based marine object recognition are given.

Robotics, computer vision, and classification tasks are just a few of the many applications of artificial neural networks. Their architecture is based on the enormously complex, parallel, nonlinear computing capabilities of the human brain. Like the brain's neurons, artificial neural networks can be programmed to carry out targeted and rapid computations, such as directing movement and vision. In this study, Galić *et al.* [18] analyze the features of real neural networks and how artificial networks might mimic these features. Two main points are presented in their paper. They begin by investigating the feasibility of using ANNs for visual identification in otherwise healthy humans. Second, they look for signs of common kidney disorders around the world and try to diagnose them. They focus on kidney cancer, kidney cysts, and polycystic kidney disease. They want to use machine learning algorithms to evaluate various samples in order to aid in the diagnosis of renal disorders.

The third part is comparison of different approaches and their effectiveness in reducing dimensionality for offline object recognition. Five stages were proposed by Elaraby *et al.* [19] to classify chest X-rays of patients infected with COVID-19. A range of machine learning techniques are used to benchmark the suggested model using different assessment metrics. Next, a backend system and mobile client app are created for an IoT platform. The backend gets the patient photographs from the client app, processes them, and applies the suggested model to diagnosis the images. Lastly, the availability, elasticity, coverage, configurability, and reliability of AWS are tapped into via a cloud-based architecture. The accuracy of 99.31% in the results beat that of other models. Therefore, real-time automatic early COVID-19 detection could be made possible by the suggested method.

A novel method combining two existing techniques is introduced for identifying salient features in high-dimensional data [20]. The data was obtained from the Alzheimer's disease neuroimaging initiative (ADNI), comprising medical images and assessments for 900 patients tracked over at least 3 years following initial examination. This new approach unites two current methods adept at solving large optimization problems with numerous data dimensions - random forests and partial swarm optimization. These were chosen for their demonstrated effectiveness in handling high dimensionality. Evaluations of the new technique exhibit superior performance over most other published approaches for this task. The accuracy of predicting Alzheimer's disease stage reached 95% across all stages of the disease. This hybrid approach leverages the strengths of two established techniques to achieve excellent results in feature selection for complex, high-dimensional medical data. Rani *et al.* [21] aim in their review paper to show how images can achieve understanding from their visual characteristics through self-supervised methods. They also explain the terminology used in self-supervised

learning and the different learning approaches such as contrastive learning and transfer learning. This review paper describes in detail the workflow of self-supervised learning, which has two main phases: pretext and downstream tasks. Near the conclusion, the authors highlight various challenges faced when working on self-supervised learning.

Rani *et al.* [21] a webcam-based system that enables users to recognize and detect facial features. Their article presents a comprehensive, systematic study evaluating classic representation learning methods on class-imbalanced datasets. They demonstrate that deeper discrimination can be learned by constructing deep networks that maintain inter-cluster differences within and between groups. MobileNet, a recently proposed CNN model, provides both offline and real-time accuracy and speed for fast, stable real-time results. This CNN architecture helps solve facial identification and recognition problems. Overall, this paper reviews different approaches and models from the literature for addressing facial recognition issues. The authors find that using more layers yields better results, and combining machine learning with multiple image datasets improves facial detection and recognition classifier performance.

All related studies mentioned above did not study the combination of dimensionality reduction PCA with fisher encoding to improve the accuracy for object recognition and detection. Therefore, this paper zeroed attention on object recognition and detection by combining dimensionality reduction PCA with fisher encoding using local SIFT features, and a support vector machine classifier.

3. METHOD

This section elaborates on. Provides an explanation of every step, in the suggested approach, which is divided into two primary phases. The initial phase involves constructing the training model, for each targeted object within a setting while the subsequent phase entails recognizing each object and detecting its location within a scene.

Figure 1 explains that during the training process firstly, extracting local SIFT features, for each object individually and within a scene. Next, fisher vector encoding will be created followed by applying a dimensionality reduction technique using PCA and finally label each feature vector to build the training model with support vector machine (SVM) as the classifier (M-SVM² multi class classification model).

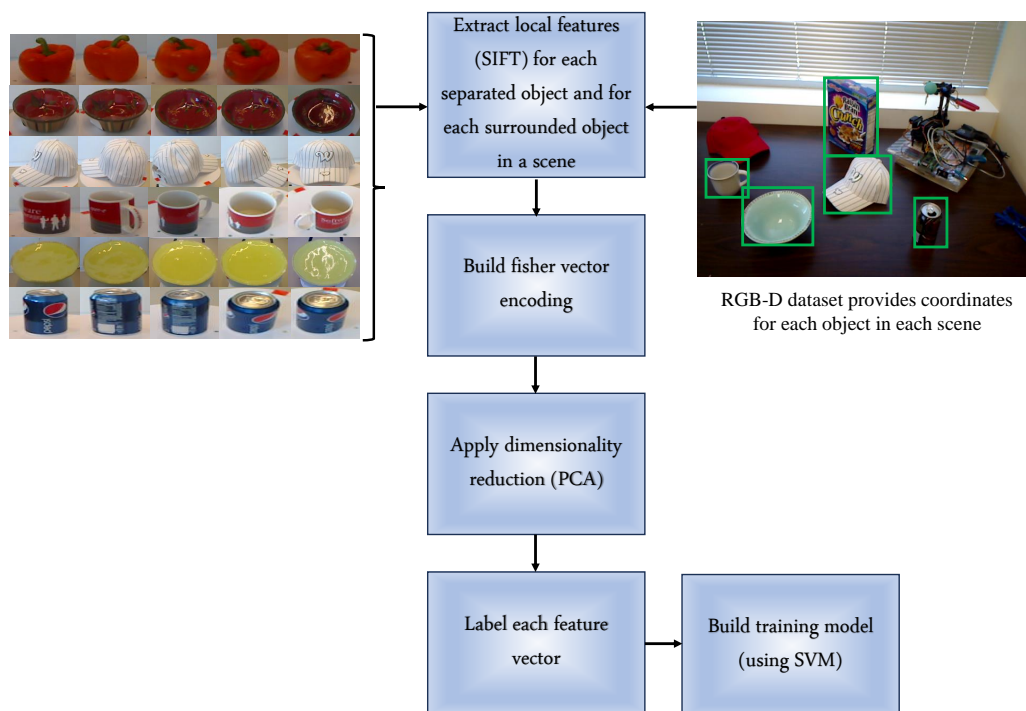


Figure 1. Training phase for object recognition and detection

While Figure 2 illustrates the testing process unfolding in two stages. Initially the first stage involves recognizing the object in isolation followed by detecting it within a scene using a sliding window technique. In the detection phase a binary SVM classifier is employed to compare the detected objects location, with the coordinates supplied by the RGB-D dataset. The following points summarize the steps in training and testing phase:

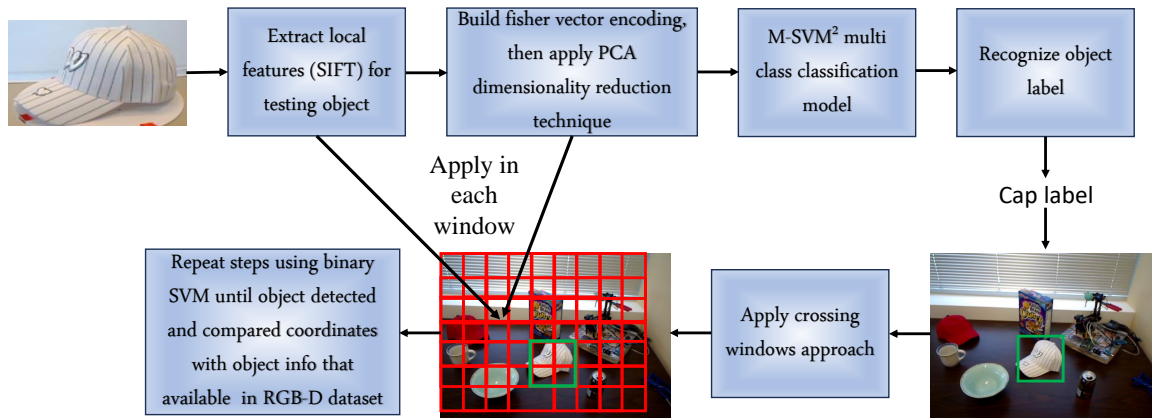


Figure 2. Testing phase for object recognition and detection

The subsections below provide an explanation and full image, for the methodology and each step in the proposed approach. The next subsections outline the recommended method, in steps. Explain the rationale behind utilizing a particular technique in the process.

3.1. Feature extraction

Feature selection is crucial, in the realm of computer vision. Thus, authors opted for employing SIFT features, known as the scale invariant feature transform, which is widely recognized as an algorithm for extracting features, in computer vision applications [22]. It was introduced by David Lowe in 1999 and has since become a cornerstone in many computer vision applications, including object recognition, image stitching, and 3D reconstruction. The key characteristic of SIFT features is their scale and rotation invariance, which makes them robust to changes in viewpoint, scale, and orientation. These features are designed to be distinctive and repeatable, enabling reliable matching and recognition across different images[23].

The SIFT feature extraction process involves several steps:

- Scale-space extrema detection: SIFT detects interest points in an image by identifying local extrema over different scales. This process involves convolving the image with a series of Gaussian filters at different scales to detect keypoints at various levels of detail. The (1) for generating the scale-space representation is as follows:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (1)$$

where $L(x, y, \sigma)$ represents the scale-space image at coordinates (x, y) and scale σ , $G(x, y, \sigma)$ is the Gaussian kernel at scale σ , and $I(x, y)$ is the input image.

- Keypoint localization: once the potential keypoints are detected, SIFT applies a detailed localization process to accurately determine their positions. It involves eliminating low-contrast keypoints and removing keypoints along edges to ensure that only stable and distinctive keypoints are retained. The refined keypoint location is determined by solving the following system of (2):

$$D(x) = \partial D / \partial x, D(y) = \partial D / \partial y, D(\sigma) = \partial D / \partial \sigma \quad (2)$$

where $D(x)$, $D(y)$, and $D(\sigma)$ are the partial derivatives of the scale-space image with respect to x , y , and σ , respectively.

- Orientation assignment: SIFT computes a dominant orientation for each keypoint to achieve rotation invariance. This step involves calculating gradient magnitudes and orientations in the local neighborhood of each keypoint and assigning a dominant orientation based on the histogram of these orientations.
- Descriptor generation: SIFT generates a descriptor for each keypoint, which encodes the local image information around the keypoint. The descriptor captures the gradient magnitudes and orientations in a region around the keypoint and represents them in a highly distinctive and compact form. This descriptor is designed to be robust to changes in scale, rotation, and illumination.

The SIFT algorithm produces a collection of keypoints, directions, and descriptors that characterize unique areas in the image [24]. These descriptors can be utilized for numerous computer vision jobs like matching keypoints across images, object recognition, or aligning images to create panoramas. SIFT features have been extensively adopted because of their robustness, repeatability, and distinctiveness, making them very useful for many computer vision tasks [25]. However, it should be noted that SIFT is a patented method, so using it may require licensing agreements in certain situations.

3.2. Feature encoding

In our method the next stage involves constructing a feature encoding vector, which is crucial, in computer vision. This process transforms raw image data into formats that are well suited for efficient processing extracting features reducing dimensionality and creating meaningful representations of the input data [26], [27]. Authors have decided use fisher encoding, where in a study [28] the researchers have demonstrated the effectiveness of using fisher vector for tasks related to recognition. This paper validates the efficiency of employing fisher encoding in recognizing objects due, to its capacity to retain information. By utilizing a GMM with M components and parameters $\{w_m, \mu_m, \sigma_m, m = 1 \dots M\}$ (weight, mean standard deviation) the fisher vector encodes local features as the gradient of their probability, in the GMM. The (3) and (4) detailing the gradient concerning the standard deviation of each component are provided.

$$G_{\mu_m} = \frac{1}{N\sqrt{w_m}} \sum_{i=1}^N p_{im} \left(\frac{x_i - \mu_m}{\sigma} \right) \quad (3)$$

$$G_{\sigma_m} = \frac{1}{N\sqrt{w_m}} \sum_{i=1}^N p_{im} \left(\frac{(x_i - \mu_m)^2}{\sigma} - 1 \right) \quad (4)$$

In the fisher vector all measurements have been combined the gradients of in each component (p_{im}). This vector has a dimensionality of 4096 which comes from multiplying the dimensionality(s) of a feature ($S = 128$) by the number of components ($M = 16$). Since this dimension is often large principal component analysis has been applied to refine the vector.

It's worth noting that the Fisher encoding technique can be computationally demanding, especially when dealing with large-scale datasets or high-dimensional feature spaces [29]. However, advancements in hardware and optimization techniques have made it more feasible for real-world applications.

3.3. Dimensionality reduction

The third stage of the proposed approach is feature encoding reduction. Where, there are several dimensionality reduction techniques that can be employed [30]. In our methodology PCA where used as a step after building fisher encoding vector. It's important to consider that the choice of dimensionality reduction method relies on the features of the data and objectives of the computer vision project. Often PCA technique is favoured over others when the aim is to enhance the distinction, between classes in tasks like classification and object recognition that require clear differentiation, between various categories [31]. Therefore, this paper compare PCA technique with others like linear discriminant analysis (LDA), and random projection (RP), please see section 4.3..

However, PCA is a widely used linear dimensionality reduction technique that aims to find a lower-dimensional representation of the data while preserving its most important information [32], [33]. It achieves this by identifying the directions (principal components) along which the data varies the most. The main steps involved in PCA are as follows [34]:

- Centring the data: to centre the data, the value will be subtracted from each feature. This process guarantees that the data is positioned at the centre point ensuring accuracy and consistency.

- Covariance matrix computation: the covariance matrix is calculated to understand how various data features are related to each other.
- Eigenvalue decomposition: the matrix of covariance is broken down into its eigenvectors and eigenvalues. The eigenvectors stand for the components while the associated eigenvalues show how much variance is clarified by each component.
- Selection of principal components: the eigenvectors are ordered according to their eigenvalues and a group of the eigenvectors (known as components) is chosen to account for the majority of variations, in the dataset. These identified components create a space, with dimensions.
- Projection: the information is mapped onto the chosen components to create a simplified representation, in dimensions.

PCA transforms a P-dimensional observed data vector y into a lower D-dimensional space. Each observation x in this compressed space is obtained by solving for the following (5):

$$x = W^T(y - \mu) \quad (5)$$

The P-dimensional vectors of the matrix W are the D dominant eigenvectors (v) of the sample covariance matrix, associated with the matrix's highest eigenvalues (λ). This achieves the desired linear transformation of the data, where W is a $P \times D$ matrix and μ is the mean of the data. PCA is an unsupervised technique, meaning it does not take class labels into account during the dimensionality reduction process. It aims to capture the overall variance in the data without considering the discriminative information specific to the object recognition task.

3.4. Supervised algorithm

The final stage of the proposed approach is building a training model that were used in the training and testing phases. In our method, the multiclass M-SVM² was used to build the model and use it for testing. Moreover, binary SVM were used to detect the place of object in a scene. The multiclass M-SVM² is a supervised classification algorithm that extends the binary SVM to handle multiclass classification problems [35], [36]. It is designed to effectively classify data points into multiple classes by constructing a set of hyperplanes that separate the classes in the feature space.

The key idea behind multiclass M-SVM² is to decompose the original multiclass problem into a set of binary sub-problems, where each sub-problem involves distinguishing one class from the rest. The M-SVM² algorithm constructs multiple binary SVM classifiers and combines their outputs to make multiclass predictions. Here is a thorough explanation of how the suggested method applied SVM techniques to develop the model, for object recognition and detection:

- Data representation: the input data for multiclass M-SVM² consists of a set of labeled samples, where each sample is represented by a feature vector and associated with a class label. The feature vectors can be derived from the RGB-D object dataset in the context of your methodology.
- Binary decomposition: the multiclass problem is decomposed into multiple binary subproblems. For N classes, N binary subproblems are created, where each subproblem focuses on separating one class from the remaining N-1 classes. This decomposition can be done using various strategies, such as one-vs-all (also called one-vs-rest) or one-vs-one. The decision function for each binary classifier can be represented as:

$$f(x) = \text{sign}(w_i^T x + b_i) \quad (6)$$

where w_i and b_i are the weight vector and bias term for the i-th binary classifier, respectively, and 'x' is the input sample.

- Binary SVM training: for each binary subproblem, a binary SVM classifier is trained. The training involves selecting the samples relevant to the specific binary subproblem and assigning appropriate labels (e.g., +1 for the target class and -1 for the other classes). The SVM algorithm learns a separating hyperplane that maximally separates the two classes in the feature space. The objective function of a binary SVM is formulated as minimizing the following expression:

$$\|w\|^2 + C \sum [\max(0, 1 - y_i(w^T x_i + b))]^2 \quad (7)$$

where $\|w\|^2$ represents the squared L2-norm of the weight vector 'w', C is the regularization parameter that controls the trade-off between the margin and the classification error, (x_i, y_i) are the training samples with inputs x_i and corresponding class labels y_i (either +1 or -1), and b is the bias term, see Figure 3.

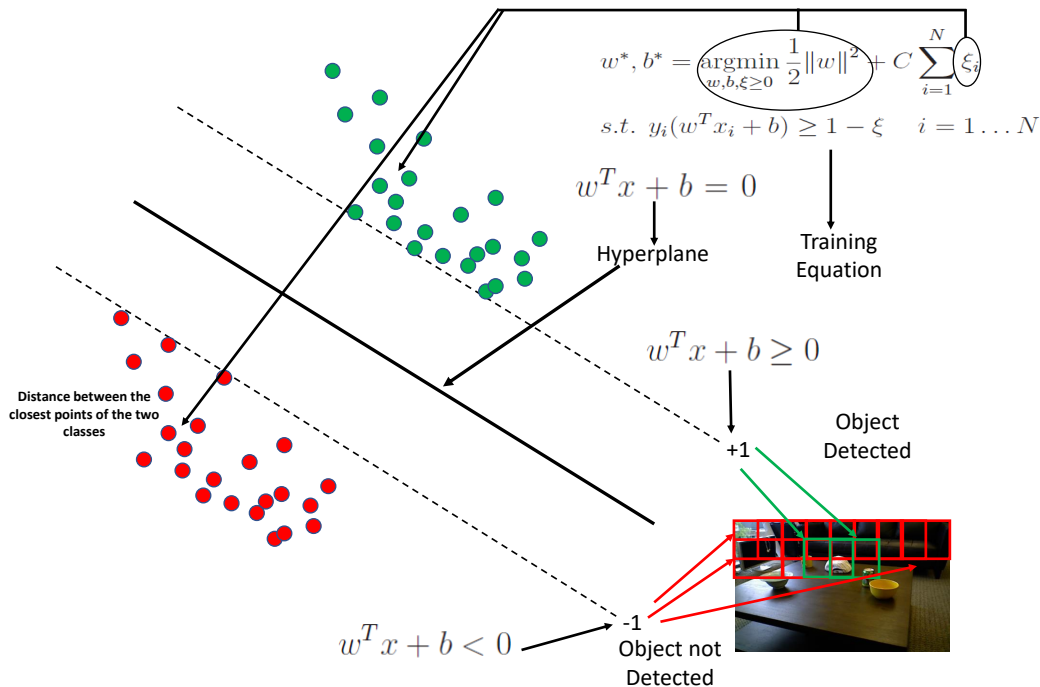


Figure 3. Visualization of binary SVM that is applied to detect object in a scene, note: this picture was used in one of our previous work [37]

- Classification decision: to make a multiclass prediction for a given test sample, the trained binary SVM classifiers are applied. Each binary classifier produces a decision score or distance that represents the sample's proximity to the corresponding class. The class label with the highest score is assigned to the test sample as the predicted class.
- Combining binary classifiers: the outputs of the binary SVM classifiers can be combined in different ways to obtain a final multiclass decision. Common methods include voting (where each classifier has equal weight), weighted voting (where classifiers are weighted based on their confidence), or using a more sophisticated fusion technique like error-correcting output codes (ECOC).

The choice of binary decomposition strategy and decision fusion method can impact the algorithm's performance and should be selected based on the specific requirements of the RGB-D object recognition task. It's worth noting that there are alternative multiclass SVM approaches, such as one-vs-one or one-vs-all SVM, as well as other multiclass classification algorithms like decision trees, random forests, or neural networks [38]. The selection of the multiclass M-SVM² algorithm depends on factors such as the nature of the dataset, computational efficiency, and the algorithm's performance on the specific problem at hand.

4. RESULTS AND DISCUSSION

This section focused on testing and evaluation of the proposed approach. It is done by browsing a comprehensive description dataset that used the experiments. Evaluation criteria employed in the suggested approach and examine the outcomes. It is concerning recognition and detection recall, precision, and F1-measure.

4.1. Dataset

Like how ImageNet categorizes things the RGB-D object dataset consists of a collection of 300 common household items grouped into 51 categories based on their similarities. To create this dataset a 3D camera

similar, to Kinect was utilized to capture depth data at a rate of 30 Hz while taking aligned RGB images at a resolution of 640x480. Each object was placed on a rotating platform. Filmed as it completed a rotation to create video sequences [39]. Three video clips were recorded for each object from camera heights to capture perspectives in relation, to the horizon [39]. In contrast to other datasets like Caltech 101 and ImageNet, which have a broad "dog" category with many different dog images that are not uniquely identifiable, the RGB-D object dataset divides objects into types and distinct instances [39]. The dataset also contains the real pose details for all 300 objects.

In this research seven objects were picked out, at random without considering their identification or spotting. The chosen items comprised of bell peppers, bowls, hats, cereal boxes, coffee cups, food plates and soda cans. Figure 4 presents a sample of separated objects used in this paper, while Figure 5 presents a sample of different scenes contains objects used in this paper. The field is rapidly evolving thanks to datasets like this one which provide an invaluable tested for developing the next generation of intelligent systems.



Figure 4. Sample of separated objects in RGB-D dataset [39]



Figure 5. sample of scenes contains objects in RGB-D dataset [39]

4.2. Evaluation

Here, explain the evaluation process in the proposed method. In this study the evaluation process comprises three components. First a training model was created using images containing objects or objects situated together in a scene. The X and Y coordinate of each object in a scenes were obtained from the RGB-D dataset. Secondly our proposed method was tested in two ways; initially each object was tested independently to determine if it was correctly identified compared to objects, without considering scenes. This testing phase utilized M-SVM² as described. The second testing approach involved generating sliding crossing windows with each scene window being assessed to ascertain if it contains the object using binary SVM. The third part showcases how the method was tested our by contrasting the strategies employed in our approach, with those utilized in studies, for recognizing and detecting objects.

Two key metrics (recall and precision) are measured in the evaluation by using true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). TP refers to correctly matching a detected object with the ground truth, while FP is incorrectly matching a detected object with the ground truth. TN is correctly not detecting an object at the current ground truth location, and FN is incorrectly not detecting an object at the current ground truth location. Recall and precision are calculated using the (8), (9):

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

In our evaluation and explanation of the significance of blending encoding and dimensionality techniques, in our method we studied object recognition and detection using SIFT feature independently once. Then, with the integration of encoding and dimensionality reduction techniques.

4.3. Experimental results and discussion

This paper presents and discusses four types of results. The initial one involves evaluating the effectiveness of the suggested method by assessing how well it identifies an object, on its own and determining its location within a scene, amidst objects as depicted in Table 1 and Table 2 measured by recall, precision and F1-score. These findings showcase the efficacy of the proposed approach whether in identifying objects or pinpointing their locations among surrounding objects, in a scene.

Table 1. The proposed approach results, recall, precision, and F1-score for object recognition individually

| Object | Recall (%) | Precision (%) | F1-score (%) |
|--------------|------------|---------------|--------------|
| Bell Pepper | 91.2 | 94.3 | 92.7 |
| Bowls | 87.5 | 88.9 | 88.2 |
| Caps | 85.1 | 86.3 | 85.7 |
| Cereal Boxes | 93.4 | 95.2 | 94.3 |
| Coffee Mugs | 89.0 | 89.7 | 89.3 |
| Food Plate | 92.6 | 94.8 | 93.7 |
| cans | 90.2 | 90.7 | 90.4 |

Table 2. The proposed approach results, recall, precision, and F1-score for object detection using different scenes

| Object | Recall (%) | Precision (%) | F1-score (%) |
|--------------|------------|---------------|--------------|
| Bell Pepper | 88.7 | 92.6 | 90.6 |
| Bowls | 80.1 | 86.5 | 83.1 |
| Caps | 81.6 | 84.5 | 83.0 |
| Cereal Boxes | 91 | 96.0 | 93.4 |
| Coffee Mugs | 88.6 | 87.9 | 88.2 |
| Food Plate | 89.6 | 91.4 | 90.4 |
| cans | 89.7 | 86.8 | 88.2 |

The second set of findings, on the side emphasized aspects of the suggested approach;

- The significance of combining dimensionality reduction methods with encoding and features for object recognition and location detection was highlighted. A comparison was made with outcomes achieved using features or using features with encoding but without incorporating dimensionality reduction techniques.
- The decision to choose the Fisher encoding method over options was explained by examining outcomes associated with encoding types such, as the VLAD encoding technique.

All the points mentioned above are detailed in Table 3 and Table 4. The outcomes were clearly depicted, showing that the suggested method performed better, than other approaches.

Table 3. Average of recall and precision for the proposed approach with different approaches

| Experiment | Approach | Recall (%) | Precision (%) |
|--------------|---|------------|---------------|
| Experiment 1 | SIFT | 74.9 | 74.9 |
| Experiment 2 | SIFT + VLAD without PCA | 78.7 | 78.8 |
| Experiment 3 | SIFT + Fisher without PCA | 80.3 | 80.4 |
| Experiment 4 | SIFT + VLAD + PCA | 83.7 | 84.0 |
| Experiment 5 | The Proposed approach (SIFT + Fisher + PCA) | 87.04 | 89.38 |

Table 4. Recall, precision, and F1-score for object detection in different scenes using only SIFT features without dimensionality reduction PCA and without encoding techniques

| Object | Recall (%) | Precision (%) | F1-score (%) |
|--------------|------------|---------------|--------------|
| Bell Pepper | 74.1 | 75.2 | 74.6 |
| Bowls | 68.8 | 69.7 | 69.2 |
| Caps | 74.8 | 73.5 | 74.1 |
| Cereal Boxes | 63.7 | 66.2 | 64.9 |
| Coffee Mugs | 82.4 | 80.1 | 81.2 |
| Food Plate | 79.3 | 77.7 | 78.4 |
| cans | 81.5 | 82.1 | 81.7 |

In the third set of results; authors explained their choice of using the PCA dimensionality reduction method of techniques, like LDA [40], and RP [41], in object detection and recognition. For more details on a comparison between PCA and other dimensionality reduction methods check Table 5. In the fourth part of the findings comparisons were made among supervised learning models used in this study. The results indicated that SVM was found to be more appropriate compared to models such, as Naive Bayes and random decision forest algorithms. These findings were visually presented in Figure 6.

Table 5. The average of recall, precision, and F1-score for object detection in different scenes using the proposed approach with different dimensionality reduction techniques

| Experiment | Dimensionality reduction technique | Recall (%) | Precision (%) |
|--------------|------------------------------------|------------|---------------|
| Experiment 1 | LDA [40] | 86.2 | 78.9 |
| Experiment 2 | RP [41] | 85.3 | 86.7 |
| Experiment 3 | PCA | 87.04 | 89.38 |

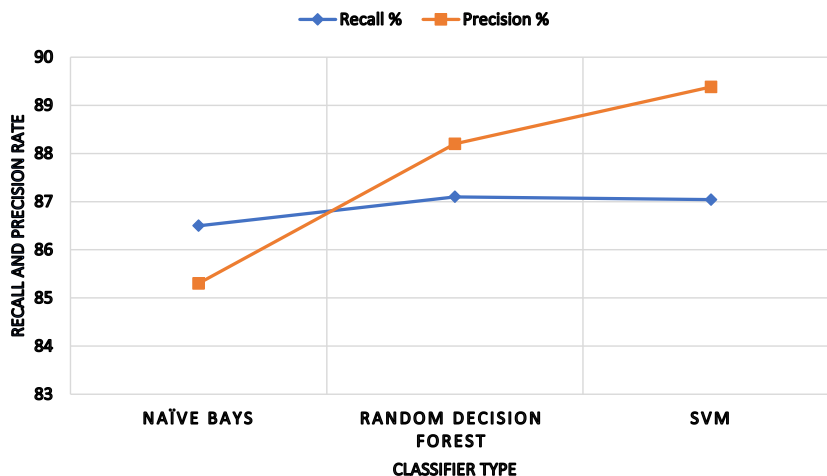


Figure 6. Results with different supervised techniques

While the proposed method has shown results, in object recognition and detection it's important to note that this doesn't always mean other techniques are inferior or that the techniques employed in the proposed

approach are always superior. The outcomes could vary based on factors like the number of objects studied or the dataset utilized. Thus methodologies in this field can vary depending on the dataset characteristics and the desired objectives. Moving forward expanding the variety of objects studied and using datasets will be crucial, in determining whether the proposed methodology consistently outperforms others.

5. CONCLUSION

This research paper introduces an approach for recognizing and detecting objects in RGBD datasets by merging fisher encoding and PCA dimensionality reduction. The study includes our experiments and the application of our method on the RGBD object dataset. The results from the experiments demonstrated that our approach yields outcomes in terms of recall, precision, F1-score and detection rate by integrating encoding with Dimensionality reduction compared to using either technique alone. Hence the experiments have shown that the proposed method (utilizing SIFT features fisher encoding, PCA, crossing sliding windows and SVM classifier) is highly practical and efficient for object recognition and detection, in RGBD environments. In work the authors plan to enhance the proposed method by incorporating a range of objects and datasets.




REFERENCES

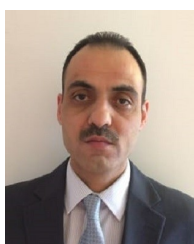
- [1] Susanto, J. A. R. Silitonga, R. Analia, E. R. Jamzuri, and D. S. Pamungkas, "Tiny-YOLO distance measurement and object detection coordination system for the BarelangFC robot," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 6, pp. 6926–6939, 2023, doi: 10.11591/ijece.v13i6.pp6926-6939.
- [2] A. Vahab, M. S. Naik, P. G. Raikar, and P. S. R., "Applications of object detection system," *International Research Journal of Engineering and Technology*, vol. 6, no. 4, p. 4186, 2008, [Online]. Available: www.irjet.net.
- [3] W. Zhiqiang and L. Jun, "A review of object detection based on convolutional neural network," in *2017 36th Chinese Control Conference (CCC)*, Jul. 2017, vol. 1157, pp. 11104–11109, doi: 10.23919/ChiCC.2017.8029130.
- [4] A. Hanafi, L. Elaachak, and M. Bouhorma, "Machine learning based augmented reality for improved learning application through object detection algorithms," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 1724–1733, 2023, doi: 10.11591/ijece.v13i2.pp1724-1733.
- [5] D. Prasad, "Survey of the problem of object detection in real images," *International Journal of Image Processing (IJIP)*, vol. 6, no. 6, pp. 441–466, 2012.
- [6] E. Maeda, "Dimensionality reduction," *Computer Vision: A Reference Guide*, pp. 1–4, 2020.
- [7] E. Postma, "Dimensionality reduction: a comparative review dimensionality reduction: a comparative review," *Journal of Machine Learning Research*, vol. 10, no. October 2016, pp. 1–35, 2007.
- [8] T. Hu, W. Wang, J. Gu, Z. Xia, J. Zhang, and B. Wang, "Research on apple object detection and localization method based on improved YOLOX and RGB-D images," *Agronomy*, vol. 13, no. 7, p. 1816, 2023, doi: 10.3390/agronomy13071816.
- [9] H. Du, S. Zhao, D. Zhang, and J. Wu, "Novel clustering-based approach for local outlier detection," in *Proceedings - IEEE INFOCOM*, 2016, vol. 2016, pp. 802–811, doi: 10.1109/INFOCOMW.2016.7562187.
- [10] T. Zhou, D. P. Fan, M. M. Cheng, J. Shen, and L. Shao, "RGB-D salient object detection: a survey," *Computational Visual Media*, vol. 7, no. 1, pp. 37–69, 2021, doi: 10.1007/s41095-020-0199-z.
- [11] N. Migenda, R. Moller, and W. Schenck, "Adaptive dimensionality reduction for neural network-based online principal component analysis," *PLoS ONE*, vol. 16, no. 3 March, p. e0248896, 2021, doi: 10.1371/journal.pone.0248896.
- [12] C. T. Nguyen, V. T. M. Khuong, H. T. Nguyen, and M. Nakagawa, "CNN based spatial classification features for clustering offline handwritten mathematical expressions," *Pattern Recognition Letters*, vol. 131, pp. 113–120, 2020, doi: 10.1016/j.patrec.2019.12.015.
- [13] M. N. Mohammad, C. U. Kumari, A. S. D. Murthy, B. O. L. Jagan, and K. Saikumar, "Implementation of online and offline product selection system using FCNN deep learning: product analysis," *Materials Today: Proceedings*, vol. 45, pp. 2171–2178, 2021, doi: 10.1016/j.matpr.2020.10.072.
- [14] A. S. Tarawneh, C. Celik, A. B. Hassanat, and D. Chetverikov, "Detailed investigation of deep features with sparse representation and dimensionality reduction in CBIR: A comparative study," *Intelligent Data Analysis*, vol. 24, no. 1, pp. 47–68, 2020, doi: 10.3233/IDA-184411.
- [15] G. Saleem, U. I. Bajwa, and R. H. Raza, "Toward human activity recognition: a survey," *Neural Computing and Applications*, vol. 35, no. 5, pp. 4145–4182, 2023, doi: 10.1007/s00521-022-07937-4.
- [16] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," *arXiv preprint arXiv:2005.04757*, 2020, [Online]. Available: http://arxiv.org/abs/2005.04757.
- [17] N. Wang, Y. Wang, and M. J. Er, "Review on deep learning techniques for marine object recognition: architectures and algorithms," *Control Engineering Practice*, vol. 118, p. 104458, 2022, doi: 10.1016/j.conengprac.2020.104458.
- [18] D. Galić, Z. Stojanović, and E. Čajić, "Application of neural networks and machine learning in image recognition," *Tehnicky Vjesnik*, vol. 31, no. 1, pp. 316–323, 2024, doi: 10.17559/TV-20230621000751.
- [19] M. E. Elaraby, A. A. Ewees, and A. M. Anter, "A robust IoT-based cloud model for COVID-19 prediction using advanced machine learning technique," *Biomedical Signal Processing and Control*, vol. 87, p. 105542, 2024, doi: 10.1016/j.bspc.2023.105542.
- [20] N. ElZawawi, H. Saber, M. Hashem, and T. Gharib, "An efficient hybrid approach for diagnosis high dimensional data for alzheimer's diseases using machine learning algorithms," *International Journal of Intelligent Computing and Information Sciences*, vol. 0, no. 0, pp. 1–15, 2022, doi: 10.21608/ijicis.2022.116420.1153.




- [21] V. Rani, S. T. Nabi, M. Kumar, A. Mittal, and K. Kumar, "Self-supervised learning: a succinct review," *Archives of Computational Methods in Engineering*, vol. 30, no. 4, pp. 2761–2775, 2023, doi: 10.1007/s11831-023-09884-2.
- [22] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.
- [23] T. Lindeberg, "Scale invariant feature transform," *Scholarpedia*, vol. 7, no. 5, p. 10491, 2012, doi: 10.4249/scholarpedia.10491.
- [24] M. Darshana and B. Asim, "Comparison of feature detection and matching approaches: SIFT and SURF," *GRD Journals- Global Research and Development Journal for Engineering*, vol. 2, no. 4, p. 7, 2017.
- [25] A. Yong and Z. Hong, "SIFT matching method based on K nearest neighbor support feature points," in *2016 IEEE International Conference on Signal and Image Processing, ICSIP 2016*, 2017, pp. 64–68, doi: 10.1109/SIPROCESS.2016.7888225.
- [26] G. Wimmer, A. Vécsei, M. Häfner, and A. Uhl, "Fisher encoding of convolutional neural network features for endoscopic image classification," *Journal of Medical Imaging*, vol. 5, no. 03, p. 1, 2018, doi: 10.1117/1.jmi.5.3.034504.
- [27] L. Liu, C. Shen, L. Wang, A. V. D. Hengel, and C. Wang, "Encoding high dimensional local features by sparse coding based fisher vectors," *Advances in Neural Information Processing Systems*, vol. 2, no. January, pp. 1143–1151, 2014.
- [28] J. Sánchez, F. Perronnin, and Z. Akata, "Fisher vectors for fine-grained visual categorization," in *Fine-Grained Visual Categorization Workshop in IEEE Computer Vision and Pattern Recognition*, 2011, p. 2010. [Online]. Available: <http://hal.archives-ouvertes.fr/hal-00817681/>.
- [29] M. Aslan, A. Sengur, Y. Xiao, H. Wang, M. C. Ince, and X. Ma, "Shape feature encoding via fisher vector for efficient fall detection in depth-videos," *Applied Soft Computing Journal*, vol. 37, pp. 1023–1028, 2015, doi: 10.1016/j.asoc.2014.12.035.
- [30] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001, doi: 10.1109/34.908974.
- [31] S. ASonawale and R. Ade, "Dimensionality reduction: an effective technique for feature selection," *International Journal of Computer Applications*, vol. 117, no. 3, pp. 18–23, 2015, doi: 10.5120/20535-2893.
- [32] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, Mar. 1993, doi: 10.1016/0098-3004(93)90090-R.
- [33] M. Verleysen and M. Verleysen, "Principal component analysis (PCA)," *Statistics*, no. September, pp. 1–8, 2001. A. Tharwat, "Principal component analysis-a tutorial," *International Journal of Applied Pattern Recognition*, vol. 3, no. 3, pp. 197–240, 2016.
- [34] A. Tharwat, "Principal component analysis-a tutorial," *International Journal of Applied Pattern Recognition*, vol. 3, no. 3, pp. 197–240, 2016.
- [35] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002, doi: 10.1109/72.991427.
- [36] Z. Wang and X. Xue, "Multi-class support vector machine," *Support vector machines applications*, pp. 23–48, 2014.
- [37] S. Awwad, B. Igried, M. Wedyan, and M. Alshira'h, "Hybrid features for object detection in RGB-D scenes," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 2, pp. 1073–1083, 2021, doi: 10.11591/ijeecs.v23.i2.pp1073-1083.
- [38] E. Mayoraz and E. Alpaydin, "Support vector machines for multi-class classification," in *International Work-Conference on Artificial Neural Networks*, 1999, pp. 833–842.
- [39] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *2011 IEEE international conference on robotics and automation*, 2011, pp. 1817–1824.
- [40] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, 2021, doi: 10.30880/jscdm.2021.02.01.003.
- [41] I. K. Fodor, "A survey of dimension reduction techniques," *Library*, vol. 18, no. 1, pp. 1–18, 2002, doi: 10.2172/15002155.

BIOGRAPHIES OF AUTHORS






Sari Awwad    is currently working as an Associate professor at the faculty of Prince Al-Hussein Bin Abdullah II for Information Technology at the Hashemite University, Jordan. In the period from 2018-2019, he has worked as the head of department of computer science and applications at the Hashemite University, Jordan. In the period from 2005-2014, he has worked as an instructor at the faculty of Prince Al-Hussein Bin Abdullah II for Information Technology. In Feb 2017, he finished his Ph.D. degree in computer science from the school of computing and communications at the University of Technology, Sydney (UTS), Australia. He has excellent research and academic skills that he gained from his research and teaching work. His current research interests include artificial intelligence applications in computer science, computer vision, machine learning, and natural language's processing. He can be contacted at email: sari@hu.edu.jo.






Ahmad Al-Rababa'a    received his bachelor's degree in computer engineering from Yarmouk University, Jordan; an M.Sc. in computer science from Amman Arab University, Jordan; and a Ph.D. in computer science from Laval University, Quebec, Canada. Ahmad currently works as an assistant professor at the Information Technology Faculty, World Islamic Sciences and Education University (WISE), Amman, Jordan. His research interests are in the areas of artificial intelligence, data compression and coding, data science, and natural language processing. He can be contacted at email: ahmad.rababaa@wise.edu.jo.






Salah Taamneh    is currently an Assistant Professor at the Department of Computer Science and Applications, Hashemite University, Zarqa, Jordan. He received the B.S. degree in computer science from Jordan University of Science and Technology, Irbid, Jordan, in 2005, the M.S. degree in computer science from Prairie View A&M University, Prairie View, Texas, in 2011 and the Ph.D. degree in computer science from University of Houston, Houston, Texas, USA, in 2016. He. His current research interests include parallel and distributed computing, machine learning and human-computer interactio. He can be contacted at email: taamneh@hu.edu.jo.



Subhieh M. El-Salhi    obtained her B.Sc. in computer science from the Hashemite University in 2000, her M.Sc. degree from the University of Jordan in 2003, and her Ph.D. degree from the University of Liverpool, Liverpool, UK in 2014. She is currently an associate professor and the chairman of the department of Computer Information System, Faculty of Prince al Hussein bin Abdullah II for Information Technology, Hashemite University. Her research interests include big data and deep learning, machine Learning and applied data mining, classification techniques, 3D surface representation, feature extraction for non-standard data sets, large data sets and statistical analysis in the context of data mining field. She can be contacted at email: subhieh@hu.edu.jo.



Ala Mughaid    was born in Irbid, Jordan, in 1984. He received the B.Sc. degree in Computer Science from Jordan University of Science and Technology (JUST), Jordan, in 2006, and the MSC in Engineering degree in engineering Computer Network from the Western Sydney University, Sydney, Australia, in 2010. Dr. Mughaid received the Ph.D. degree in Computer Science from Newcastle University – Sydney, Australia, in 2018. In 2018, Dr. Mughaid joined the Department of Information Technology, The Hashemite University, as an assistant professor, Zarqa, Jordan. Dr. Mughaid current research interests include but not limited to cyber security, cloud computing, image processing, artificial intelligence, virtual reality, and data mining. He is working voluntarily in many social services. He can be contacted at email: ala.mughaid@hu.edu.jo.