Date of publication JUN-30, 2025, date of current version MAY-12, 2025.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.24.2.4020

A Hybrid CNN-Multiclass SVM Model for Household Object Recognition: A System for Domestic Robotic Vision

SMITA GOUR[1], PUSHPA B. PATIL[2]

[1]Department of Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot-587102, Karnataka, India
[2]Department of Computer Science and Engineering (Data Science), BLDEA's V P Dr P G Halakatti College of Engineering & Technology, Vijayapur-586102, Karnataka, India

Corresponding author: Smita Gour (e-mail: smita.gour@gmail.com)

ABSTRACT Computer vision is a broad area covering many aspects of detection and recognition of objects. Household object recognition is a challenging problem, mainly due to lighting conditions, differences in shape, size of objects, position, occlusion, clutter of the objects and background. A novel deep learning and multiclass Support Vector Machine (SVM) based mechanism for household object recognition, is presented here. A comprehensive dataset of household objects, Washington RGB-D, is used. The Convolutional Neural Network (CNN) is used for feature extraction. It involves building and training CNN model using RGB-D dataset. Once CNN model is trained the features are automatically extracted from fully connected layer. The normalized feature vectors are used as input to train the SVM classifier. Empirical analysis of the proposed system's performance is carried out using RGB-D and real-time datasets. This hybrid model classifies the real-time instances with an accuracy of 89%. It exhibited an accuracy of 90.2% with RGB-D standard dataset, taken alone. The robustness and accuracy of the proposed approach pave the way for improved interaction between robots and humans in interior environments.

KEYWORDS Deep Learning; Convolution Neural Network; Tensorflow; Object Recognition; RGB-D dataset

I. INTRODUCTION

Computer vision is affected by illumination from different types of sources, distance between objects and the camera, background clutter and the like. Computer vision includes image acquisition, preprocessing, analysis in finding symbolic decisions. Computer vision is a demanding field for robotics, attempting object recognition. Service robots are used for identifying and classifying the objects in domestic environment. These robots need to recognize the objects, move to the location of the object and work in a constrained environment. Hence the work was proposed. The task may be carried out either in off-line mode, or in real-time mode (the object in the site of a camera). But recognition of object from an image is more complex than object in online mode, mainly due to blurred images, noise, occlusion, clutter and illumination conditions. Object recognition mechanisms can rely on different aspects, as follows.

CAD like primitives: Use basic structures such as edges, primal sketch, etc.

Structural parts of object: Use structural parts.

Basic appearance of object: Template matching

Feature based mechanisms: Invariant features

This paper introduces a comprehensive object recognition system designed to empower robots to perceive and recognize diverse objects within their surroundings. Leveraging the capabilities of Convolutional Neural Networks (CNNs), theproposed system aims to address the intricacies associated with extracting robust and accurate features of household objects. Support Vector Machines (SVMs) are combined with CNN to classify household objects such as cup, book, plate, spoon, chair etc. Literature survey is carried out to know the state-of-the-art in the area.

In literature many methods that attempt to recognize the objects from the image are reported. Some of them are discussed here.This survey has been done considering two categories such as traditional Machine Learning (ML) approaches and modern deep learning approaches.

To build an efficient household object recognition system, an analysis of datasets [1] of household objects is needed for AI-enabled techniques. A new dataset to estimate the six-degrees of freedom (DoF) pose estimation of known objects is presented [4]. A pose evaluation metric ADD-H based on the Hungarian assignment algorithm is used [6]. Introduce 2 mainstream detection frameworks, incorporate 5 major object

detection challenges and related solutions, list 4 well-known On-Board Diagnosis (OBD) datasets and OBD evaluation metrics, and list some applications related to OBD. Also a comprehensive study [2] on generic object detection is done through which issues in general object recognition are identified [2]. Also presents Convolution Neural Network (CNN) and different Deep Convolution Neural Network architectures, along with the known datasets and definitive metrics. Many research works based on popular traditional object detection methods [3] such as SVM, KNN and like are exists along with the current trends in deep learning-based methods [3] to solve the problem of object recognition.

Under tradition ML approach [5] is one of approach where visual information is extracted and used to recognize objects. They propose a method that uses multiple features like color and shape features. The proposed method incorporates the results from each classifier (k-nearest neighbor, kNN) using simple probabilistic methods to get strong recognition results for household objects. The results show that method can get (84.02) ± 18.85 % for household object recognition under extreme conditions.Another method [8] and [20] uses the Point Cloud Library (PCL), a three-stage mechanism to address the object recognition challenge. To prepare the object images for feature extraction, segmentation techniques are applied to it in the first step of PCL. Using a segmented image, appropriate shape-based features that can distinguish between different types of objects are extracted in the second stage. The final step uses a support vector machine to classify and recognize the object of that particular kind (SVM). The accuracy rate for the system for ten distinct sorts of household objects is 91%. In [16] The Histograms of Oriented Gradients (HOG) technique is used to extract features from images. Principal Component Analysis (PCA) has been applied on the extracted features. For classification of objects Support Vector Machine (SVM), Random Forest, Input mapped classifier, Multi-Layer Perceptron (MLP) classifier and Gaussian process classifier have been applied. [19] This method includes extracting shape and texture features from the object images, removing the shadow that separates the object from its shadow. Then objects are categorized to their respective classes using Back Propagation Neural Network (BPNN). The accuracy range provided by the system is 81%-92% for 10-25 distinct object kinds [14]. Presents a realistic deep learning based mechanism for home object recognition. Three or more layers are used in its implementation, each of which extracts one or more image features. RAS, a Robot activity Support system [15] is presented. It connects physical robots with smart environment technologies to assist with daily tasks. This study focuses on how numerous components, such as map building, object detection, navigation, user interface, and activity error detection. Overall results show that Support Vector Machines (SVM) based on Principle Component Analysis (PCA) performs better than other methods by showing accuracy of 92.02% and highest F-Score measurement. The combination of Support Vector Machine (SVM) and Histogram of Gradient (HoG) is presented [17]. It is used in object identification to find suspicious objects. Additionally, it demonstrates that the system can identify an intruder with 89% accuracy.

Under modern deep learning approach [7] Compare three feed forward Deep Convolutional Neural Networks (DCNNs) with human observers (N = 45). Furthermore, they present evidence that, when compared to a typical ResNet architecture, a DCNN named vNet, which has biologically plausible receptive field sizes, demonstrates superior agreement with human viewing behavior. In order to provide a unique control input for a computer vision-based robot arm, [9] proposed a work which describes the design and evaluation of a real-time framework that combines speech recognition, camera-based object detection, and an inference module [11]. Here, the DCNN-based object detection algorithms are presented in parts like backbone networks, loss functions and training strategies; classical object detection architectures, complex problems, datasets and evaluation metrics, applications and future development directions. Scorbot-ER 5 Plus robotic arm [12] for the pick and place task of household goods is presented. Prior to grasping, a deep learning approach based on AlexNet is used to identify the object's class. The findings of the experimental analysis show that the AlexNet-based technique performs well in terms of detection and categorization of the grabbed items, showing good accuracy. Another convolution neural network MobileNetV2-based model [13] is developed for classifying and detecting common household tools. First, MobileNetV2 is chosen to serve as the feature extraction backbone network. Then, the full connection layer network and Softmax classifier are used to realize the classification and recognition of common household tools. Faster R-CNN (Regional Convolution Neural Network) [18] is another classifier to handle static object identification and detection with low precision. In [22], they introduce a straightforward and scalable detection algorithm that enhances the mean Average Precision (mAP) by over 30%. Their method builds on two key principles: (1) leveraging high-capacity CNNs applied to bottom-up region proposals for effective object localization and segmentation, and (2) utilizing supervised pre-training on an auxiliary task followed by domain-specific fine-tuning to address the challenge of limited labeled training data, resulting in a significant performance improvement. [23] Present a unified framework of kernel features for depth images that effectively model size, 3D shape, and depth edges. Through comprehensive object recognition experiments, author demonstrate that: local features capture diverse and complementary cues from a single depth frame or view and their approach significantly enhances the performance of both depth and RGB-D (color + depth) recognition, achieving a 10–15% accuracy improvement over the current state-of-the-art methods. In [24], they propose a model that combines CNNs and Recursive Neural Networks (RNNs) to effectively learn features and classify RGB-D images [25]. Presents Hierarchical Matching Pursuit (HMP) for RGB-D data, an unsupervised method that leverages sparse coding to learn hierarchical feature representations directly from raw RGB-D inputs. Extensive experiments across multiple datasets demonstrate that the features learned through HMP achieve state-of-the-art object recognition performance when combined with linear support vector machines. The approach [10] merges multiple Gaussian Mixture Models (GMM) using a probabilistic approach. This system is capable of capturing household objects with dimensions greater than 3 cm × 3 cm × 2 cm and weighing less than 800 g. Furthermore, with an F1-score of 76.43%, the suggested systems are able to extract 40 items, each of which encompassed 40 poses.

Through this survey it is found that there are some limitations in traditional ML approaches such as dependence

on handcrafted features, poor performance in cluttered or noisy scenes and scalability issues. Although modern deep learning techniques have significantly solved these problems, their application in real-world home environments remains constrained by several limitations. A major concern is their reliance on extensive, annotated datasets, which are often expensive and labor-intensive to create—particularly when accounting for the wide variability of household objects in different domestic contexts. Furthermore, these models demand considerable computational resources for both training and inference, making them less suitable for integration into low-power, resource-constrained platforms commonly used in home service robots. So it has been decided to build a hybrid model which incorporates the advantages of both ML and DL approaches. It is concluded that SVM is highly accurate and used as classifier and deep learning can extracts powerful features for household object recognition. Hence SVM and CNN are combined and proposed this hybrid model.

II. PROPOSED METHODOLOGY

The proposed methodology shown in Fig. 1 consists of three phases. These are 1) Building and training CNN model for feature extraction. 2) building and training SVM model using these features to classify household objects. 3) Testing phase. CNN Model will be trained using RGB-D dataset and knowledge base is created which will be fed to SVM model to train it to recognize objects. Each of these blocks is discussed in this section.

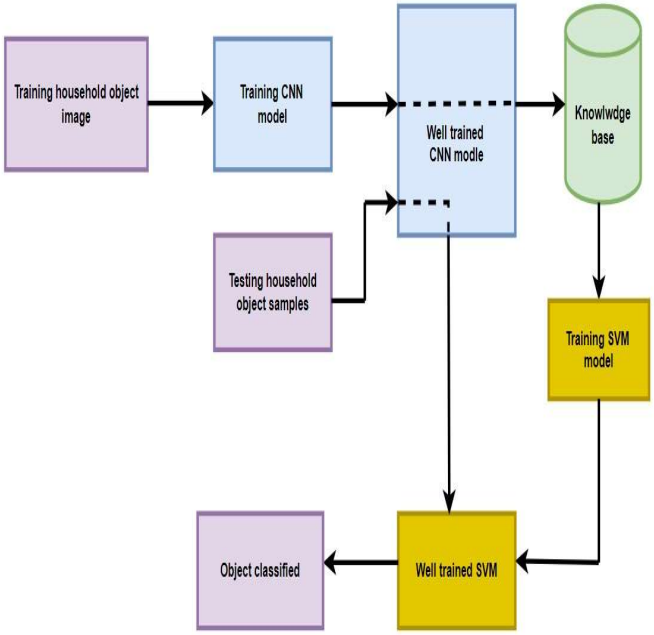

Figure 1. Proposed Block Diagram

A. WASHINGTON RGB-D DATASET

The standard Washington RGB-D data set [21] has been used to evaluate the model initially, since most of the works on object recognition are using this dataset. The RGB-D Object Dataset is a large dataset of 300 common household objects.This dataset is built using a Kinect 3D camera that records contemporise and aligned RGB and depth images of size 640x480 at 30 Hz. Each object was placed on a rotating table and for each rotation sequences of video were captured. The dataset also provides ground truth pose information for all 300 objects. Some of example Objects from RGB-D object dataset are shown in Fig. 2.

Figure 2. Samples From RGB-D Dataset

Some of real time object images shown in Fig. 3 are also captured using webcam, like cell phone, watch, water bottle, spectacle, bag, bucket, mug, exam pad, toothbrush and pen. On this set, augmentation is applied to create more images to well train CNN model. This process is discussed in following sections.

Figure 3. Captured Images

B. DATA AUGMENTATION

Data augmentation is the process of creating much number of images by changing the appearance or the global view of existing images. It helps to regularize and hence avoid over fitting while training a model.

The imagedatagenerator function is used for this process. It creates many image batches and it will apply various transformations randomly on them like rotating, flipping, shipping and shearing. At the end we get eventually are many more images in each of these batches that will help in training. This augmentation process requires some parameters to be set such as range of shear which alter image by stretching it in some direction, range of zoom which increase the size of image, vertical and horizontal flips to toggle the image, rescaling the values of pixels, filling gaps, rotating images and shifting width and height of the image. Table 1 shows all these important parameters and their range of values used in this work during data augmentation. Once data is ready the CNN model is to be created. To create and train CNN model some parameters and their values must be selected. This we call it as hyperparameter tuning which is discussed in following section.

Table 1.Tensorflow Data Augmentation Parameters

Sl. No.	Parameter	Value
1	Range of shear	0.4
2	Range of zoom	0.2
3	Horizontal flip	True
4	Vertical flip	True
5	Rescale	1/255
6	Mode of filling	'nearest'
7	Angle of rotation	25
8	Width shift	0.1
9	Height shift	0.1

E. HYPERPARAMETER TUNING

Initially we often cannot decide the architecture of optimal model and thus it is necessary to explore a range of possibilities of different architectures which mainly differ by some parameters. These parameters define the model architecture are referred to as hyperparameters and the process of searching ideal values of these parameters for the model to be optimal is referred as parameter tuning. In true machine learning fashion, this hyper parameter tuning can be automatic. Some of the hyper parameters and their corresponding values considered in this work are shown in Table 2.

Table 2.CNN Parameters and their values

Parameter	Why?	Values
Number_of_layers	The total number of layers in the CNN, including convolutional, pooling, and fully connected layers. More layers allow the network to learn more complex features. (e.g., 5, 10, 20)	16
filters	The number of filters in a convolutional layer. Each filter can learn a different feature. (e.g., 32, 64, 128,256,512)	Final 512
kernel_size	Size of the convolutional filters. Smaller kernels capture fine details, larger kernels capture broader features. (e.g., (3, 3), (5, 5))	(3,3)
stride	The step size for moving the filter across the input image. Larger strides reduce the spatial dimensions faster. (e.g., (1, 1), (2, 2))	(2,2)
padding	Determines whether the input volume is padded with zeros around the border. 'same' keeps spatial dimensions, 'valid' reduces them.	'valid'
activation	Indicate activation function to be used. Controls the output of each neuron. (e.g., relu, sigmoid, tanh, softmax)	'relu'
pool_size	The size of the window for pooling operations, which reduce the dimensionality of the feature maps. (e.g., (2, 2))	(2,2)
dropout_rate	Fraction of input units to drop during training. Prevents overfitting by randomly turning off neurons. (20% to 50%)	20%
batch_size	The number of samples per batch during training. Larger batch sizes can lead to more stable gradients but require more memory. (e.g., 32, 64, 128)	32
epochs	Number of training iterations over the entire dataset. More epochs can improve accuracy but may cause overfitting. (e.g., 10, 50, 100)	50
learning_rate	Step size at each iteration while moving toward a minimum of the loss function. (e.g., 0.001, 0.01)	0.01

With these all parameters the model has been built to extract features for recognizing household objects which is discussed in further sections. For the task of household object recognition, the 5-fold cross validation using 10 object dataset has been carried out to get optimal values for the two parameters such asnumber of filters and number of epochs. The result of these cross validations are shown in Table 3 and Table 4 respectively.

Table 3.Results of cross validation on number of epochs

Number of Epochs	Accuracy
10	0.89
20	0.91
30	0.91
40	0.93
50	0.94

Table 4. Results of cross validation on number of filters

Number of Filters	Accuracy
32	0.94
64	0.941
128	0.95
256	0.955
512	0.96

G. CNN MODEL CREATION FOR FEATURE EXTRACTION

The most commonly used deep learning technique is Convolutional Neural Network (CNN). A CNN convolutes features from input data using 2D convolution layers. CNN does the job of extraction of features by itself. Hence, there is no need of identifying and extracting features manually. These models are highly prescribed for the tasks of computer vision, like object classification. The convolutional neural network has four layers. They are convolution, pooling, flattening and fully connected layer.

We assume a gray scale or binary image 'I' represented by size n_1 x n_2 defined by a function given in Eq. (1).

$$I: \{1,2 \dots n_1\} \times \{1,2 \dots n_2\} \rightarrow W \subseteq \mathrm{R} \qquad (1)$$

Given the filter $K \in R_2 h_1 + 1$, the discrete convolution of the image I with filter K is given by Eq. (2).

$$\mathrm{I} * \mathrm{K} = \sum_{u=-h1}^{h1} \sum_{v=-h2}^{h2} \mathrm{K}_{u,v} \mathrm{I}_{r+u,s+v} \quad (2)$$

Where the filter K is given by Eq. (3)

$$\mathrm{K} = \begin{pmatrix} \mathrm{k}_{-h1,-h2} & \cdots & \mathrm{k}_{-h1,h2} \\ \vdots & \mathrm{k}_{0,0} & \vdots \\ \mathrm{k}_{h1,-h2} & \cdots & \mathrm{k}_{h1,h2} \end{pmatrix} \qquad (3)$$

Discrete Gaussian filter $K_{G\{\sigma\}}$ which is given by Eq. (4), is used for smoothening.

$$\mathrm{K}_{G(\sigma)} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp 2^{\left(\frac{r^2+s^2}{2\sigma^2}\right)} \qquad (4)$$

All these layers are implemented using Keras API which is written in Python and is capable of running on top

of TensorFlow, Microsoft Computational Network Toolkit (CNTK), or Theano. Keras incorporates basic languages like TensorFlow that enables to build deep learning models. The data preparation operations like preprocessing, augmentation for training and testing is also carried out by Tensorflow.

Tensorflow framework has been used to create the CNN model. It consists of totally 16 layers employing operations such as repeated convolution, pooling, flattening, and fully connected dense layer. Algorithm 3.1 depicts the modelbuilding. Input to this algorithm is dataset containing household object images to train the model. On these images pre-processing has been done including augmentation and reshaping of the images. Then CNN model is built based on the hyper parameter values and the dataset (step 3 to step 6). After pooling, Spatial Pyramid Pooling (SPP) has been applied on pooled features that enhanceCNN by creating fixed-size feature representations from images of varying sizes. It preserves spatial information, improves scale invariance and reducing overfitting. Thus enhanced feature extraction and robustness in object recognition tasks.

Algorithm 3.1: Proposed CNN Model
Input: Dataset consisting of images from the standard Washington RGB-D dataset
Output: CNN Features
Begin

Step 1: Data Augmentation with shear_range of 0.4 and with both horizontal and vertical flip
Step 2: Input layer:
The layer in this step performs the reshaping of data. The output of this step is square root
of a number of pixels in a given image.
Step 3: Convolution and Pooling layers
P=32
for i=1 to 5 do
Add two convolution layers with P feature maps of size of 3×3 and a rectifier activation function. Also set dropout layer at 20% in between these two layers.
Add Max Pool layer with size 2×2 and stride 2
P=P*2
Step 4: Apply SPP on pooled features
Step 5: Flatten layer:
It is going to flatten the pooled feature map into a column. Set dropout layer at 20%.
Step 6: Dense layer
Create fully connected layer with 512 units and a rectifier activation function. Also set dropout layer at 20%.
Step 7: Output: Take features out from dense layer

The overall architecture of the model built using this algorithm, to extract the features, for household object recognition, is shown in Fig. 4. These features are used to build and train SVM.

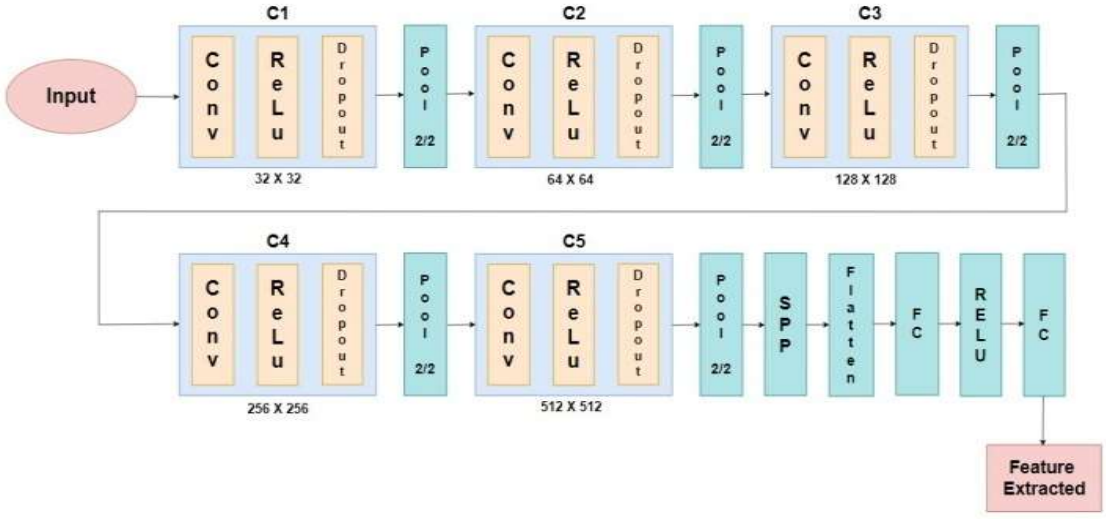

Figure 4. Proposed CNN Model

H. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for both classification and regression tasks. SVM works by finding the optimal hyperplane that best separates the data into different classes. By carefully preprocessing the data, extracting meaningful features, and selecting appropriate kernel functions, SVMs can achieve high accuracy and generalize well to new, unseen objects. This makes them suitable for various applications, including automated home systems, robotics, and smart appliances.

a) SVM Kernels

For household object recognition tasks using SVM, the choice of kernel can significantly influence the model's performance. Here are some suitable kernels commonly used for such tasks.

Linear Kernel: It is used for datasets where objects can be separated with a straight line. It's computationally efficient and often used when the number of features is large compared to the number of samples. It is given by (5).

$$K(x, y) = x.y \qquad (5)$$

Polynomial Kernel: Suitable for capturing interactions between features, especially if the relationship between household objects is polynomial in nature. It is depicted using the (6).

$$K(x, y) = (\gamma x.y + r)^d, \qquad (6)$$

where D is a Degree, r is coefficient and γ is scaling factor.

Radial Basis Function (RBF) Kernel: Highly effective for non-linear data where the distinction between household objects is not clear-cut. It can handle varying shapes and sizes of objects well using (7).

$$K(x, y) = e^{(-\gamma\|x-y\|^2)} \quad (7)$$

where, γ is scaling factor

Sigmoid Kernel: As shown in (8) useful for certain non-linear problems. It can be used when household objects have complex and non-linear separations.

$$K(x, y) = \tanh(\gamma x.\, y + r) \quad (8)$$

where, γ is scaling factor and r is coefficient

b) Different Types of Features

Along with choice of kernel functions, different categories of features are also considered in the experimentation to show how powerful the CNN features are. They are as follows.

Color and Shape features: Color and shape features are essential for object recognition in robotic vision systems, providing complementary information to differentiate household items. Color features capture the visual characteristics of objects through color spaces like RGB, Hue Saturation Value(HSV), and Lab, which are resilient to various lighting conditions and environmental factors. Techniques such as color histograms, color moments, and dominant color extraction quantify the color distribution, while combining these with texture patterns enhances the feature set. In contrast, shape features emphasize the geometric structure of objects, using methods like contour-based descriptors (e.g., perimeter, circularity), region-based moments (e.g., Hu or Zernike moments), and keypoint detection for scale and rotation invariance. Advanced methods like skeletonization and Fourier descriptors provide compact, simplified representations of complex shapes. By integrating both color and shape features, robotic systems become more robust, addressing challenges like lighting variations, similar colors, and complex object shapes. This fusion ensures precise and reliable object recognition in domestic settings.

Histogram of Oriented Gradients (HOG) Features: HOG is a widely used feature descriptor in household object recognition for robotic vision systems, as it effectively captures the shape and structure of objects by analyzing gradient information in an image. The process begins by calculating the gradient of pixel intensities, highlighting edges and contours that define the object's shape. These gradients are then grouped into orientation bins within small cells (e.g., 8x8 pixels), with each bin representing the frequency of gradient directions. To ensure robustness to changes in lighting and contrast, the HOG features are normalized over larger spatial regions called blocks. The final HOG descriptor is a high-dimensional feature vector formed by concatenating the histograms from all blocks, summarizing the gradient information for object classification. In household object recognition, HOG is particularly effective because it is resilient to lighting variations, small translations, and minor rotations, making it ideal for dynamic home environments. Additionally, HOG focuses on the edges and contours of objects, which are crucial for identifying common household items such as chairs, cups, and books. Despite its efficiency, HOG may struggle with large deformations or highly complex shapes, but when combined with machine learning models like Support Vector Machines (SVMs), it offers an efficient and reliable approach for real-time object recognition in domestic robotics.

SIFT & SURF Features: SIFT (Scale-Invariant Feature Transform) and SURF (Speeded Up Robust Features) are both powerful feature extraction techniques widely used in household object recognition for robotic vision systems. These methods focus on identifying distinctive key points or interest points in an image, which are stable under various transformations such as scale, rotation, and partial occlusions—common challenges in real-world household environments. SIFT detects key points at different scales and orientations by analyzing image gradients, allowing the system to identify objects even when they appear at different sizes or rotated. SURF, a faster variant of SIFT, uses a similar approach but with optimizations for speed, making it suitable for real-time applications. It employs a "Hessian matrix" for detecting key points, which helps it perform faster than SIFT while maintaining robustness to changes in lighting, rotation, and scale. Both SIFT and SURF create descriptors based on the local appearance around key points, which are then used to match corresponding points across different images. This ability to identify and match key points enables recognition of household objects regardless of viewpoint or minor changes in object orientation. These features are particularly useful in household environments, where objects may appear from different angles or under varying lighting conditions. While SIFT and SURF are highly effective for object recognition, they can be computationally expensive, but their robustness and ability to handle object variations make them valuable tools in real-time robotic vision systems for identifying household items.

For household object recognition, the RBF kernel is often a preferred choice due to its flexibility in handling non-linear relationships. However, depending on the specific characteristics of the dataset and the computational resources available, other kernels like the Polynomial and Sigmoid kernels might also be considered. Experimentation and cross-validation are typically used (shown in Table 5) to determine the best kernel for a given task. Considering all the kernels with the same validation set and different set of features, the accuracy of the model has been obtained shown in Table 5. It depicts that RBF has given better results for household object recognition than other kernels.

Table5.Experimentation with different combination of kernels and features

Features / Kernel Type	Color and Shape	HOG Features	SIFT & SURF Features	CNN Features (Proposed)
Linear	63.5	71.2	76	81
Polynomial	62.1	71	70	88
Sigmoid	60.3	69	68	83
RBF (Proposed)	78	78.8	79	**90.2**

c) Radial Basis Function (RBF) Kernel

RBF kernel can handle complex and varied patterns found in household object images, capturing non-linear relationships effectively. It generally provides better performance in high-dimensional spaces, which is common in image data after feature extraction. With proper parameter tuning, RBF can achieve a good balance between bias and variance, leading to robust model performance. RBF kernal is formulated as in (9).

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \qquad (9)$$

where, $\left\| x_i - x_j \right\|^2$ is the squared Euclidean distance between the two points x_i and x_j and γ (gamma) is a parameter that defines the spread of the kernel. It is a positive real number. This parameter controls the width of the Gaussian function. A smaller γ value means a broader Gaussian (less sensitive to individual data points), while a larger γ value means a narrower Gaussian (more sensitive to individual data points). The choice of γ is crucial for the performance of the SVM with the RBF kernel

d) The SVM Objective Function

The SVM optimization problem aims to find the hyperplane (or decision boundary) that maximizes the margin between different classes while allowing some misclassifications to prevent overfitting. For a given training set (x_i, y_i) where x_i are the feature vectors and y_i are the corresponding class labels, the SVM optimization problem with the RBF kernel can be formulated as in (10).

$$\min_{w,b,s} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} s_i \qquad (10)$$

subject to the constraints:

$$y_{i(w\cdot\phi(x_i)+b)\geq 1-s_i} \ \forall_i$$
$$s_i \geq 0 \ \forall_i$$

where:w is the weight vector.
b is the bias term.
s_i slack variables that allow for misclassifications.
$\phi(x_i)$ is the feature transformation applied by the RBF kernel

The regularization parameter C controls the trade-off between achieving a large margin and minimizing the classification error on the training data. With higher C value, the model aims to classify all training examples correctly, which can lead to a smaller margin and potential overfitting, if the data is noisy.

By carefully tuning both C and γ using methods like grid search with cross-validation, a well-performing SVM model can be created for household object recognition tasks.

e) Grid search with cross-validation

Grid search with cross-validation is an effective method for tuning hyperparameters like C and γ in SVMs with an RBF kernel. It ensures that the chosen parameters provide the best possible performance while avoiding overfitting and underfitting. In this study a grid search is performed over a range of C and γ values and cross validation of 10-fold to optimize the SVM model's performance in recognizing household objects. The experiment involved varying C from a small value 0.001 to a large value 100 and γ from values like 0.001 to 10 and the accuracy obtained with each combination is shown in Table 6. The results in Table 6 typically show that as C increases, the model becomes more complex, potentially overfitting the data, while lower γ values correspond to broader decision boundaries that can generalize better across different objects. Finally, the optimal values found for C and γ for RGB-D object recognition are 10 and 0.1 respectively with the highest accuracy of 90.2%.

Table 6.Result of grid search and cross validation.

C \ λ	0.00	0.01	0.1	1	10
0.001	60.1	64.7	69	73	68
0.01	63.2	67.8	72.4	79	71.3
0.1	66.4	70.2	74.2	82	78.4
1	66.9	76.3	80.1	85	80.1
10	72	78	88	**90.2**	86
100	70	69.4	66.6	64.2	60

III. RESULTS AND DISCUSSION

As discussed in the previous sections different experimentations have been conducted to build an acceptable model. That is the CNN model for feature extraction is tuned with the parameters like data augmentation parameters along with number of epochs, filters etc. SVM model for recognition of objects is tuned with parameters like kernel functions, C and λ. The overall tuned model for household object recognition is depicted in Table 7.

The model trained and evaluated using RGB-D dataset, with tuned parameters, shown in Table 4. This model is evaluated against other existing methods. The results are validated with both the standard dataset i.e., Washington RGB-D data set as well as real time dataset. Comparison with other existing methods is presented in the next section.

Table 7.Final parameters of proposed model with RGB-D dataset

RGB-D dataset		CNN Parameters (Tuned)			SVM Parametrs (Tuned)			
No. of objects	No. of training samples	No. of epochs	No. of filters	Data Augmentation	C	λ	Kernel	Accuracy(%)
50	80X50=4000	50	512	Yes/With shear range 0.4 and with both horizontal and vertical flip	10	1	RBF	90.2

A. COMPARISON WITH EXISTING SYSTEMS

Many articles related to object recognition based on machine learning exist in literature. Those using RGB-D data set are considered for comparison with the proposed model, for accuracy (shown in Table 8). In the first method, texture histograms and color histograms are utilized for RGB feature extraction, and spin images and SIFT descriptors are employed for depth feature extraction. For classification, the linear SVM is employed [21]. In second method, RGB and depth features are extracted using a mixture of hierarchical

kernel descriptors, and the classification is done using a linear SVM [22]. For feature extraction, a collection of kernel descriptors is employed, and for classification, linear SVM is employed in third method [23]. Fourth method [24] utilizes a model based on a CNN and RNN combo. The method [25] uses raw RGB-D data, the HMP approach is utilized to automatically learn hierarchical feature representations. The method in [26] is advanced deep learning approach that is a transformer based CNN. It capture richer contextual information. The proposed model uses combined CNN features with SVM classifier and the accuracy obtained is acceptable as compared to other methods in Table 8. The same is depicted using the graph showed in Fig.5.

Table 8. Performance Comparison with Existing Model

Reference	Key elements	Accuracy (%)
[21]	SVM	81.9
[22]	KDES	84.1
[23]	Kernel Descriptor	86.2
[24]	CNN-RNN	86.2
[25]	RGB-DHMP	88.5
[5]	Color+shape+KNN	85
[26]	CNN+TransNet	92
	Proposed Hybrid CNN-SVM Model	**90.2**

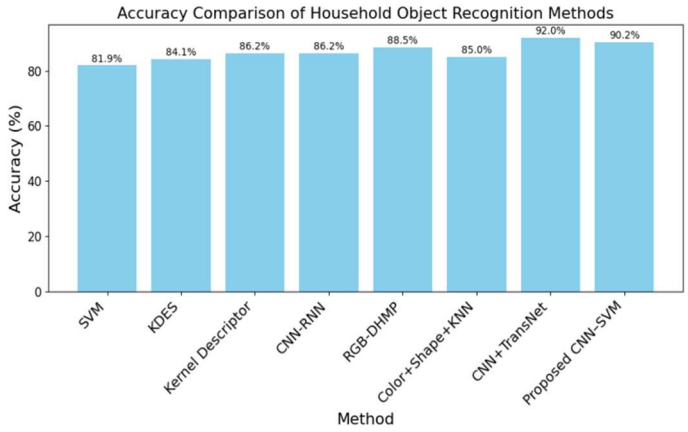

Figure 5. Graph on comparisons

The proposed CNN–SVM hybrid model achieves an accuracy of approximately 90.2%, effectively bridging the gap between traditional machine learning and deep learning approaches. In this architecture, a Convolutional Neural Network (CNN) is employed to extract high-level features from input images, while a Support Vector Machine (SVM) serves as the final classifier. This integration leverages the feature representation capabilities of CNNs alongside the robust classification margin properties of SVMs.

Previous research has shown that CNN–SVM models frequently outperform CNNs using softmax classifiers, particularly when working with smaller datasets. In this work, by utilizing a pretrained CNN and substituting the softmax layer with an SVM, the model not only improves accuracy but also reduces computational overhead. The achieved accuracy of 90.2% surpasses that of earlier non-deep methods such as standalone SVM, KDES, and KNN, and even outperforms certain pure CNN variants.

Although slightly below the highest-performing deep learning method (CNN+TransNet, 92.0%), the hybrid model demonstrates strong performance with a simpler architecture. The remaining performance gap may be attributed to the absence of transformer modules or multi-modal fusion, which enable CNN+TransNet to capture richer contextual information. Nonetheless, the CNN–SVM model proves to be a practical and effective solution, especially in scenarios with limited training data or computational constraints.

B. EXPERIMENTATION WITH REAL TIME HOUSEHOLD OBJECT IMAGES

Some of the real time household objects like Bottle, Cup, Knife, Spoon, Apple, Cell phone, Clock, Scissor, Tooth brush, Fork from the dataset shown in Fig. 2 are considered to evaluate the system performance. In this evaluation step, 50 samples of each are tested, and the confusion matrix for the same is depicted in Fig. 6. An average accuracy of 89.1% is obtained.

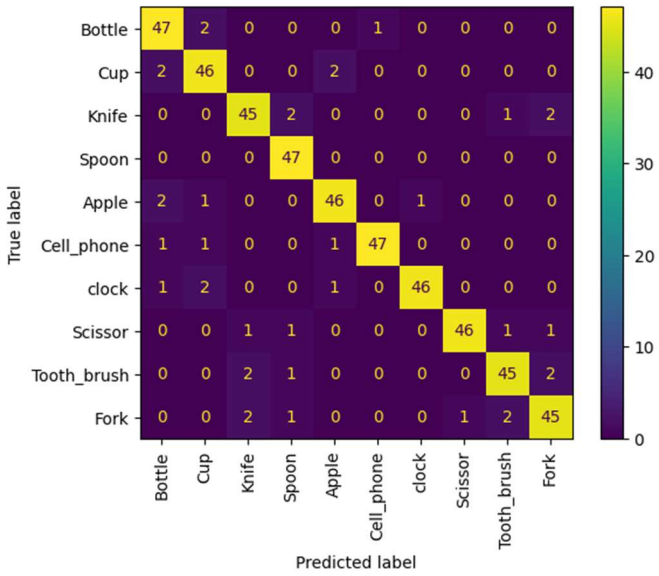

Figure 6. Confusion Matrix

IV. CONCLUSIONS

Object recognition through Computer Vision has many applications in different domains. In this work, a computer-based model is trained for identification,separation and recognition of objects from images and scene.It employs the CNN and SVM classifier.A drawback of CNN is that itrequires large amounts of input for training and with increase in the size of the data, it is observed to be slow. In this work, the emphasis is on recognizing single object in an image. The accuracy of 89.1% and 90.2%, respectively, are observed for the real-world object dataset and RGB-D dataset. This is suitable for embedding in a robot working in restricted workspaces. Limitation of this model is, the images with single objects were considered and evaluated the model. The work can be extended by building a model that recognizes multiple objects in images. Also depth information can also be utilized to build a model which may improve the performance of current model.

References

[1] D.A.Bhayana, O.P.Verma, “Analysis of existing datasets of household objects for AI-enabled techniques,” In: Sharma, H., Shrivastava, V., Bharti, K.K., Wang, L. (eds) *Communication and Intelligent Systems.*

ICCIS 2022. Lecture Notes in Networks and Systems, 2023, vol 686. Springer, Singapore. https://doi.org/10.1007/978-981-99-2100-3_4.

[2] E. Arulprakash, M. Aruldoss, "A study on generic object detection with emphasis on future research directions,"*Journal of King Saud University - Computer and Information Sciences*, vol. 34, issue 9, pp. 7347-7365, 2022.https://doi.org/10.1016/j.jksuci.2021.08.001.

[3] D. Raj, "Recent object detection techniques: a survey," *International Journal of Image, Graphics and Signal Processing*, vol. 13, no. 2, 47, 2022.https://doi.org/10.5815/ijigsp.2022.02.05.

[4] S. Tyreeet al., "6-DoF pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark," *Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, 2022, pp. 13081-13088, https://doi.org/10.1109/IROS47612.2022.9981838.

[5] M. Attamimi, D. Purwanto and R. Dikairono, "Integration of color and shape features for household object recognition," *Proceedings of the2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Semarang, Indonesia, 2021, pp. 169-174, https://doi.org/10.23919/EECSI53397.2021.96242543919/EECSI53397.2021.9624254.

[6] R. B.Wang, D. K.Yan, L.Xu, "A brief survey on recent advances of object detection with deep learning," In: Pan, JS., Li, J., Namsrai, OE., Meng, Z., Savić, M. (eds) *Advances in Intelligent Information Hiding and Multimedia Signal Processing. Smart Innovation, Systems and Technologies*, vol 211, 2021. Springer, Singapore. https://doi.org/10.1007/978-981-33-6420-2_43.

[7] L. E. Van Dyck, R. Kwitt, S. J. Denzler and W. R. Gruber, "Comparing object recognition in humans and deep convolutional neural networks – An eye tracking study,"*Front. Neurosci.*,vol. 15, 750639, 2021. https://doi.org/10.3389/fnins.2021.750639.

[8] S.Gour, P.B.Patil, B.S.Malapur, "Multi-class support vector machine-based household object recognition system using features supported by point cloud library," In: Sharma, H., Saraswat, M., Kumar, S., Bansal, J.C. (eds) *Intelligent Learning for Computer Vision. CIS 2020. Lecture Notes on Data Engineering and Communications Technologies*, 2021, vol 61. Springer, Singapore. https://doi.org/10.1007/978-981-33-4582-9_8.

[9] R. Bagwe, R. Natharani, K. George and A. Panangadan, "Natural language controlled real-time object recognition framework for household robot,"*Proceedings of the2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, NV, USA, 2021, pp. 1215-1220, https://doi.org/10.1109/CCWC51732.2021.9376044.

[10] M. Attamimi, K. Liusiani, A. N. Irfansyah, D. Purwanto and R. Dikairono, "Development of visual data acquisition systems of household objects,"*Proceedings of the2021 International Electronics Symposium (IES)*, Surabaya, Indonesia, 2021, pp. 411-416, https://doi.org/10.1109/IES53407.2021.9594002.

[11] Y.Xiao, Z.Tian, J.Yu, et al., "A review of object detection based on deep learning," *Multimedia Tools and Applications*, pp. 23729–23791, 2020. https://doi.org/10.1007/s11042-020-08976-6.

[12] M. Abbas, J. Narayan, S. Banerjee and S. K. Dwivedy, "AlexNet based real-time detection and segregation of household objects using scorbot,"*Proceedings of the2020 4th International Conference on Computational Intelligence and Networks (CINE)*, Kolkata, India, 2020, pp. 1-6, https://doi.org/10.1109/CINE48825.2020.234392.

[13] W. Chen, W. Chen, C. He, N. Liu, P. Wu and H. Shi, "Research on classification and detection system of common household tools for home service robot,"*Proceedings of the2020 International Conference*

[14] *on System Science and Engineering (ICSSE)*, Kagawa, Japan, 2020, pp. 1-5, https://doi.org/10.1109/ICSSE50014.2020.9219314.

[15] S. Gour, and P. B. Patil, "An exploration of deep learning in recognizing household objects," *Grenze International Journal of Engineering & Technology (GIJET)*, no. 6, 2020.

[16] G. Wilson, C. Pereyda, et. all, "Robot-enabled support of daily activities in smart home environments,"*Cognitive Systems Research*, vol. 54, pp. 258-272, 2019, https://doi.org/10.1016/j.cogsys.2018.10.032.

[17] Md G. Sarowar, Md. A. Razzak, and Md. A. Al Fuad, "HOG feature descriptor based PCA with SVM for efficient & accurate classification of objects in image,"*Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing (IACC)*, 2019.https://doi.org/10.1109/IACC48062.2019.8971585.

[18] N. Surantha, W. R. Wicaksono, "Design of smart home security system using object recognition and PIR sensor,"*Procedia Computer Science*, vol. 135, pp. 465-472, 2018, https://doi.org/10.1016/j.procs.2018.08.198.

[19] A. Sarkale, et al., "An innovative machine learning approach for object detection and recognition,"*Proceedings of the 2018 Second IEEE International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018. https://doi.org/10.1109/ICICCT.2018.8473221.

[20] S. Gour, and P. B. Patil, "A novel machine learning approach to recognize household objects,"*Proceedings of the 2016 IEEE International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 2016.https://doi.org/10.1109/SCOPES.2016.7955543.

[21] K. Alhamzi, M. Elmogy, S. Barakat, "3D object recognition based on local and global features using point cloud library," *International Journal of Advancements in Computing Technology (IJACT)*, vol. 7, no. 3, pp.43-54, 2015.

[22] K.Lai, L.Bo, X.Ren, D.Fox, "A large-scale hierarchical multi-view RGB-D object dataset,"*Proceedings of the IEEE International Conference on Robotics and Automation*, Shanghai, China, 9–13 May 2011, pp. 1817–1824.https://doi.org/10.1109/ICRA.2011.5980382.

[23] R.Girshick, J.Donahue, T.Darrell, J.Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation,"*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus*, OH, USA, 23–28 June 2014, pp. 580–587.https://doi.org/10.1109/CVPR.2014.81.

[24] L.Bo, X.Ren, D.Fox, "Depth kernel descriptors for object recognition,"*Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Francisco, CA, USA, 25–30 September 2011, pp. 821–826.https://doi.org/10.1109/IROS.2011.6095119.

[25] R.Socher, B.Huval, B.Bhat, C.D.Manning, A.Y.Ng, "Convolutional-recursive deep learning for 3D object classification,"*Proceedings of the International Conference on Neural Information Processing Systems*, Lake Tahoe, NV, USA, 3–6 December 2012, pp. 656–664.

[26] L.Bo, X.Ren, D.Fox, "Unsupervised feature learning for RGB-D based object recognition,"*Proceedings of the International Symposium on Experimental Robotics*, Québec City, QC, Canada, 18-21 June 2012, pp. 387–402.https://doi.org/10.1007/978-3-319-00065-7_27.

[26] Y. Zhang, M. Yin, H. Wang and C. Hua, "Cross-level multi-modal features learning with transformer for RGB-D object recognition,"*IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7121-7130, 2023, https://doi.org/10.1109/TCSVT.2023.3275814.

SMITA GOUR: *From last 15 years, Prof. Smita Gour is working as assistant professor in the Department of Computer Science and Engineering at Basaveshwar Engineering College, Bagalkote, Karnataka, India. She completed post-Graduation from Vishweshwaraya Technological University, Belagavi, Karnataka. Her field of interest is Digital Image Processing and Mahine Learning. She attended lots of National and International conference and numbers of research papers published in her field.*

PUSHPA B. PATIL: *Dr. Pushpa B. Patil, Working as Professor and Head in the Department of Computer Science and Engineering at BLDEA's CET, Vijayapur, Karnataka. She completed post-Graduation from Vishweshwaraya Technological University, Belagavi, Karnataka and PhD from Swami Ramanand Teerth Marathwada Univer-University, Nanded-431606, Maharashtra State, India. Her field of interest is Digital Image Processing and Mahine Learning. She attended lots of National and International conference and numbers of research papers published in her field.*