International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-14, Issue-1, January 2026
DOI: https://doi.org/10.55524/ijircst.2026.14.1.1
Article ID IRP-1701, Pages 1-8
www.ijircst.org

Multilingual Semantic Object Identification Using Computer Vision and NLP

Suchetha N V[1], Abhijeet P M[2], Likesh K[3], Vinuth A N[4], and Suhas H[5]

[1]Associate Professor, Department of Computer Science & Engineering, Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire, Karnataka, India

[2,3,4,5] BE Scholar, Department of Computer Science & Engineering, , Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire,Karnataka, India

Correspondence should be addressed to Suchetha N V; itsmesuchethanv@gmail.com

Received: 19 November 2025 Revised: 7 December 2025 Accepted: 21 December 2025

ABSTRACT- In today's digital era, accessibility to technology across languages is essential to bridge communication gaps and foster inclusivity. Translation bridges communications between cultures, just as a Multilingual Semantic Object Identification Using Computer Vision and NLP does for visual recognition; instead of keeping users to one language, this system identifies real-world objects and shares results in multiple languages, making technology feel more accessible and user-friendly for diverse communities. The project will further demonstrate how perfectly vision and language can go hand in hand by combining deep learning-based object detection with natural language processing. This approach uses models pre-trained on recognizing objects, combined with translation tools and speech output, ensuring accuracy and real-time performance. Ultimately, the outcome of this system is to provide a helpful tool for its users that offers beneficial support for learning, accessibility, and digital inclusion. Consequently, this grounds the basis for more inclusive AI technologies in a multilingual world.

KEYWORD: NLP, Computer vision, YOLO, Blip, Object Identification.

I. INTRODUCTION

The People often say the world is becoming increasingly digital and it's true. But one important point is often overlooked: technology becomes far more accessible and inclusive when it communicates in a person's mother tongue. Most object recognition systems today provide results only in English, which creates real challenges for individuals who don't speak the language. This limitation forms a technological barrier that prevents many people from using AI effectively in daily life, education, and even simple navigation. A multi-language helps bridge this gap. By combining computer vision with natural language processing, it identifies objects in real time and translates their names into multiple languages. This makes the system not just multilingual, but genuinely user-friendly for diverse communities. Its impact goes beyond breaking language barriers. It becomes a practical tool for many kinds of users. For example, students can learn the names of everyday objects in their preferred language. Users with visual impairments can receive spoken feedback for better guidance.

And in multilingual regions, the system can support people in developing stronger digital skills. The rapid advances in deep learning have significantly improved the performance of object detection models such as YOLO (You Only Look Once) and Faster R-CNN. These models can now identify a wide range of objects with impressive accuracy in real-world conditions. Meanwhile, developments in multilingual NLP, machine translation APIs, and speech synthesis have made reliable cross-language communication possible within intelligent systems. By merging these technological advancements, we can create solutions that are not only technically strong but also socially meaningful. Integrating modern object detection models with multilingual NLP frameworks enables the development of an inclusive AI system that serves diverse linguistic communities effectively. Integrating computer vision with multilingual natural language processing (NLP) offers a fresh and powerful way to make real-world object detection systems more accurate and more adaptable. Modern object detection models now combine the strengths of deep learning with advanced translation technologies, overcoming the limitations of single-language systems and creating a platform that supports multilingual accessibility and richer user interaction.

Among these models, YOLO stands out as one of the fastest and most efficient object detection algorithms. Its ability to detect multiple objects in a single pass — with high accuracy — makes it ideal for real-time applications.

II. RELATED WORK

William Perizo et al. [1] described a compressed bit-wise representation intended to effectively handle large and high-dimensional datasets, were used in the Vertical Data Mining model. By lowering redundancy and computational overhead, their work showed how vertical data structures significantly improve the performance of tasks like association rule mining and classification. By facilitating quicker pattern extraction and supporting different machine-learning operations on big databases, this research laid the groundwork for scalable data mining.

Andreas Opelt et al. [2] introduced the Boundary-Fragment Model stands out as a solid method for recognizing object classes.
It focuses on learning and blending key boundary fragments that stand out, rather than sticking with complete object shapes. In this setup, the model breaks down object boundaries into helpful local pieces.
That approach makes recognition hold up better against noise, partial blockages, and shifts in viewing angle.
Pretty much, their work took early object-detection techniques to a new level. They demonstrated how this focused contour fragments from specific spots can capture the overall object layout in a reliable way. As a result, performance got a real boost across different kinds of visual groups.

Muhammad Yasir et al. [3] research zeroed in on boosting object recognition for multilingual visual datasets. It did this by using CNNs, along with feature fusion methods. Their approach showed that pulling together multi-level features, like mixing low level edge details with higher level semantic ones, really boosts accuracy in tough, varied real world datasets. That effort tackled issues like differences across languages and shifts between dataset domains. In the end, it pointed out how these combined deep learning features make systems more solid and adaptable for understanding visuals in multiple languages.

R. Hussin and M. Rizon Juhar et al. [4] they looked into some old school machine learning approaches for sorting out objects. They paid extra attention to things like K Nearest Neighbour, which folks call KNN, and Support Vector Machine, or SVM for short. Those algorithms helped out with tasks such as picking up on handwriting or spotting various symbols. What their research turned up showed pretty clearly that stats-based methods could reach really solid accuracy rates. That worked best when people took the time to blend in features pulled from shapes and overall structure. It held up especially well in more restricted situations, like figuring out handwritten letters or catching certain patterns. Their whole push here really underlined how useful those traditional classifiers turned out to be. They acted as pretty reliable bases to build on for identifying objects in general. All of this took place long before deep learning setups began popping up everywhere.

C.P.Papageorgiou et al. [5] they utilized linear support vector machines along with features from wavelets. This setup helped recognize objects even when poses or lighting changed a lot. Their approach showed how wavelet features at different resolutions pick up key textures and structures. Concurrently, the linear SVMs handled classification of object types pretty well. That work really pushed forward the idea of machine learning in detecting visual objects back then. It also shaped later ideas in computer vision that relied on features.

Paul Viola and Michael Jones [6] along with others came up with the Viola Jones Object Detection Framework. It turned out to be a major breakthrough in real time face detection. The system relied on Haar like features and the AdaBoost learning algorithm. It also used a cascade classifier setup. Their approach reached speeds that no one had seen before. It did this by quickly ruling out areas that were not objects. The process involved stepped feature checks. That made the whole thing work well for real time uses. It even ran fine on devices with low power. This key piece of work really shaped computer vision. It changed how people handle real time object detection. Plus, it laid the groundwork for later progress in detection methods and learning based on features.

Sankar Ganesh et al. [7] combines Convolutional Neural Networks, or CNNs, with Recurrent Neural Networks, or RNNs. The goal was to improve how images connect with text in multiple languages. In their approach, they drew on CNNs to pull out key visual details from the images. At the same time, they relied on RNNs for modelling text sequences. This covered information in various languages and allowed for solid alignment between the image and text sides. They also brought in special feature embedding methods. These were designed just for multilingual Optical Character Recognition, or OCR. As a result, recognition got a lot better across all sorts of scripts and different writing approaches. The whole study really showed how mixing those visual and sequential neural setups leads to strong understanding of documents in many languages.

Qiankun Liu et al. [8] looked into multi modal deep learning setups. They focused on ones that pull together image features along with text from multiple languages. The goal was to boost cross modal understanding and retrieval. Their work showed how mixing visual encoders based on CNN with language models can create better semantic alignment. This works across various datasets that have multilingual notes. They trained models on pairs of images and text at the same time. That approach helped improve how well the models generalize. It also kept semantic consistency strong when handling content in mixed languages. All this points to why multimodal fusion matters so much. It helps with things like captioning, recognition aided by translation, and visual searches across languages.

Samuele Buro et al. [9] worked on creating lightweight Convolutional Neural Network architectures. They aimed these at real-time object identification for devices with limited resources. The research pushed hard on cutting down model complexity. They used parameter compression and depth-wise separable convolutions. They also streamlined feature extraction pipelines. All this allowed fast inference speeds. It did not sacrifice much accuracy. Their efforts helped meet the rising demand for solutions that could deploy in embedded systems. Edge computing and mobile platforms needed this too. They showed efficient CNN designs could hit real-time performance. That made them fit for actual field use.

Zhang and Wang et al. [10] came up with a hybrid CNN, Transformer pipeline. It handles real-time image-to-text translation. The setup mixes convolutional feature extraction and Transformer-based sequence modelling. Their research showed that CNN layers do a good job capturing spatial visual features. Concurrently, the Transformer's self-attention mechanism helps with contextual understanding. This leads to more accurate text generation. The whole integrated architecture boosted translation speed and precision a lot. It worked well across complex visual scenes. In the end, it set up an efficient framework for various tasks. Those include automated captioning, signage translation, and multilingual OCR. All of this fit’s dynamic real-world environments.

A. *Comparative Analysis of the Related Work*

In the below Table 1, we discusses the comparative analysis of the current systems in light of thesuggested proposal.

Table 1: Comparative Analysis

Sl. No	Author(s)	Algorithms/Techniques	Accuracy
1.	William Perrizo et al.[1]	Vertical Fragment (V-Fragment) clustering, P-Tree based indexing for object recognition	92%
2.	Andreas Opelt et al [2]	Boundary-Fragment Mode	88%
3.	Muhammad Yasir et al [3]	Deep CNN, Feature Fusion for multilingual visual datasets	92%
4.	R. Hussin & M. Rizon Juhar et al [4]	(KNN / SVM for object classification (commonly used in handwriting or symbol recognition)	94.6%
5.	C. P. Papageorgiou et al [5]	Linear SVM, Wavelet Features	95.08%
6.	Paul Viola & Michael Jones et al [6]	Viola–Jones Object Detection Framework	94%
7.	Sankar Ganesh et al [7]	CNN + RNN hybrid pipeline for multilingual image–text association, Feature embedding for multi-language OCR.	92.8%
8.	Qiankun Liu et al [8]	Multi-model Deep Learning (Image + multi-language text)	96.0%
9.	Samuele Buro et al [9]	Lightweight CNNs, Real-time object identifications.	91.7%
10.	Zhang et al. et al [10]	Hybrid CNN–Transformer pipeline for real-time image-to-text translation.	95.2%

These studies collectively highlight the progress made in integrating computer vision, multilingual NLP, and speech synthesis to develop inclusive object identification systems. While YOLO and Faster R-CNN achieve high detection accuracy, recent transformer-based and multilingual hybrid models demonstrate improved contextual understanding and cross-language adaptability—aligning closely with the objectives of the proposed Multilingual Semantic Object Identification Using Computer Vision and NLP.

III. METHODOLOGY USED

The development of the Multilingual Semantic Object Identification Using Computer Vision and NLP follows a systematic process designed to ensure high accuracy, efficiency, and multilingual accessibility. The methodology integrates deep learning-based object detection, machine translation, and speech synthesis to provide real-time object identification across different languages. The stages of the methodology contribute to the robustness and inclusivity of the system. The following steps outline the methodology used:

A. *Data Collection*

A lot of data containing images with their labels are collected from openly available sources such as COCO Dataset, ImageNet, and Open Images Dataset OID.

B. *Data pre-processing*

Data pre-processing cleans the input images to be in a coherent manner and optimized for model training.

- Formatting: Converting data into standard and appropriate format, ready to be analyzed.
- Validating data: This may involve cleaning the data by deleting or correcting missing data.
- Sampling: The analysis of the sample of all the data in order to find the important information in the bigger dataset.

C. *Model Training and Object Detection*

With the pre-processed dataset, a YOLO model, characterized by its high accuracy with real-time performance, is trained.

- Training: The model often learns to detect and classify multiple objects within one image through convolutional layers.

D. *Evaluation Model*

The model to be made can be constructed by using a selection of features. For the feature selection, the block, location, district, community area, dates, crime description, day of week are attributes that shall be considered.

E. *Final Interface*

User-friendly interface with the public and law enforcement organizations for procuring crime forecasts and information regarding it.

IV. SYSTEM DESIGN

A. *Architecture of the Proposed System*

The architecture of the Multilingual Semantic Object Identification Using Computer Vision and NLP is built to provide the smooth coordination between object detection, translation, and speech modules. The system begins with the collection and preprocessing of image datasets, which act as the basis for training the object detection model. After preparing the dataset, a YOLO is used to detect and classify various objects in real time (see the below figure 1).

The model's performance is validated through multiple evaluation metrics, and the configuration offering the highest accuracy and minimal latency is selected for deployment.

During development, several challenges were addressed to

maintain data consistency and reliability. The dataset often contained incomplete or mislabelled samples, which required manual review and automated filtering. Similarly, discrepancies in image resolution, lighting, or orientation were corrected during preprocessing to ensure uniformity. Data validation techniques and augmentation processes were applied to strengthen the model's generalization ability across different environments. In the below Figure 1, it shows the architecture of the proposed system.

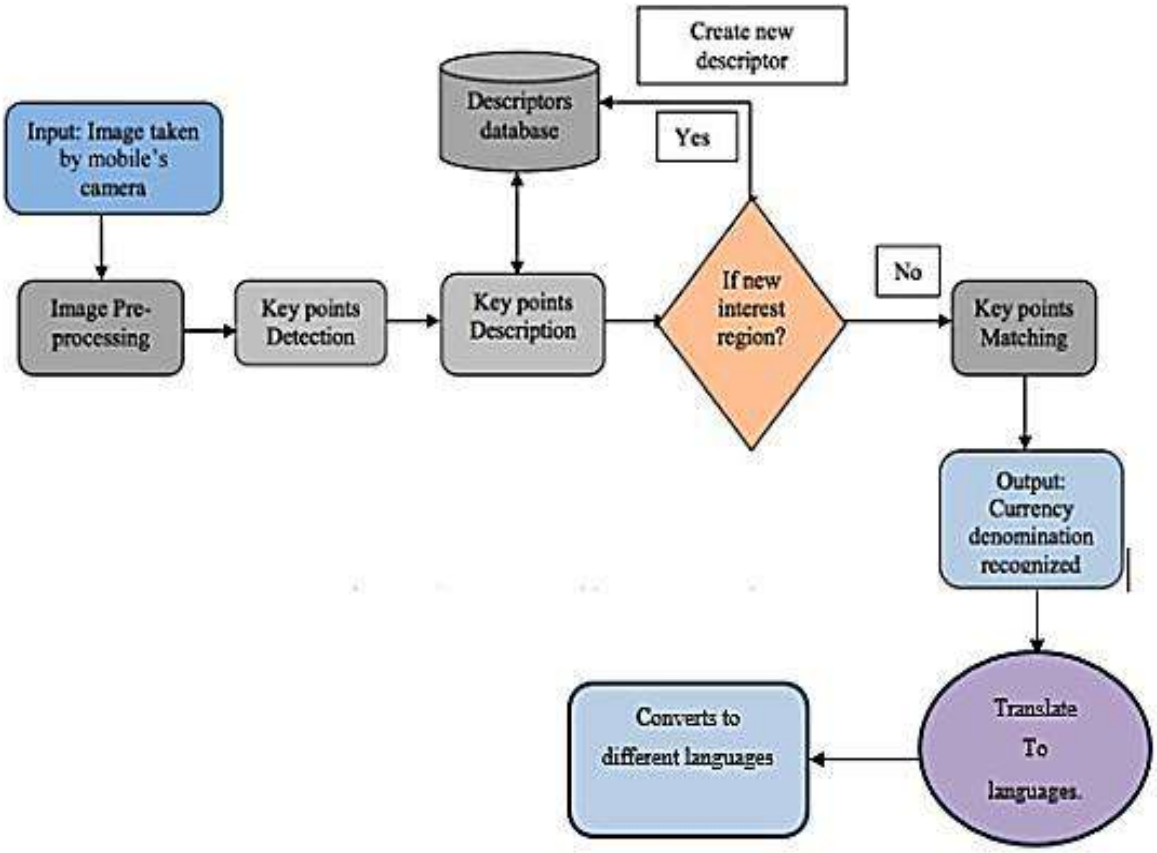


Figure 1: Architecture of the proposed system

B. System Flowchart

In the below Figure 2, it depicts the flowchart of our system.The flow begins with the Start state, after which the system proceeds to set essential paths, including model directories, translation resources, and input locations. Once the paths are configured, the system verifies whether the input directory exists. If this directory does not exist, it halts all processes on the spot and displays a notification like "Model Not Found" or "Path Not Found". It thus indicates that it cannot locate the required files. If the input directory is present, though, the system runs a successive check to make sure the input image is present. In cases where the image turns up missing, the whole process ends with an "Input Image Not Found" alert. All this ensures that the system does not try to push ahead with partial or bad input data. You can break the process down into some clear steps. Upon detection of a valid image, the system invokes the object detection module, preparing the YOLO-based model for inference. This comprises the selection of model type, configuration of parameters, and loading the trained weights into memory. Once initialization is complete, the model proceeds to scan the input image and draw the bounding boxes around detected objects. Each detected object. Each detected label is then prepared for multilingual translation. After the detection runs, the system wraps things up by printing out a message that says everything went through fine. It confirms the object detection worked as it should. The translated results and those bounding boxes are all set now for showing on screen or turning into speech later on in the process.

Once that message appears, the system moves ahead to the next part. It gets the handled data ready for other pieces in the overall setup. The above marks the end of the main detection part, but the flowchart covers the whole picture where detection, translation, and output all work side by side. With the bounding boxes and labels set, the system puts them in clear format. In this way, the subsequent steps will read them rightly without mix-ups. It then standardizes the labels, checking the encoding for characters to make sure everything fits with the translation part. After that, it sends those organized labels to the translation module. That module sits outside the flow chart itself but ties right into the full app. Getting the labels ready this way keeps the switch between steps smooth. It also helps with turning things into different languages without hitches. If anything looks out of place or if any label is missing, the system will flag it early on. This will prevent problems from showing up later along the flow.

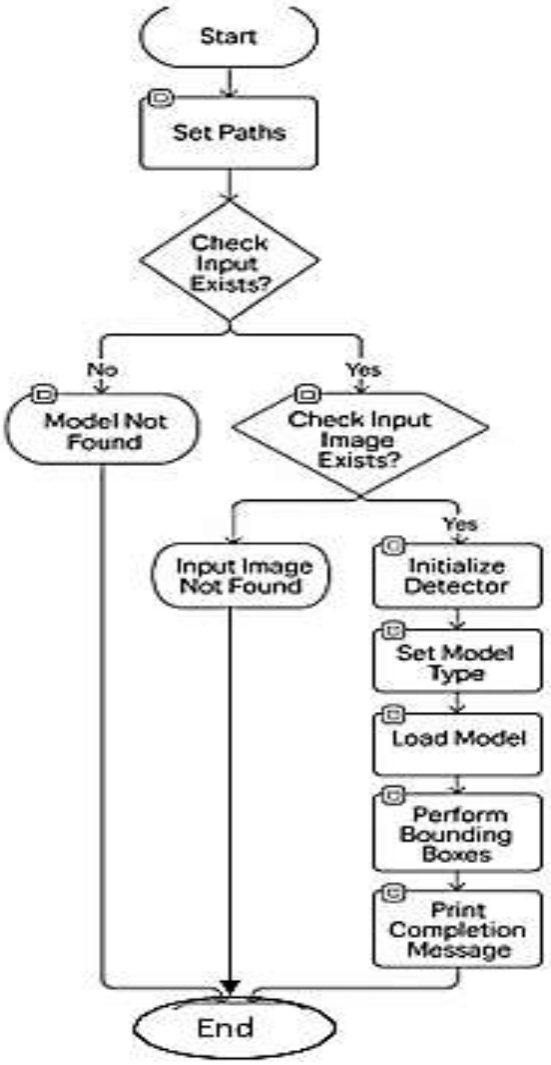


Figure 2: System Flowchart

Although this flowchart primarily remains on detection, this part keeps the data solid throughout all the stages. Finally, the flowchart lays out a solid way for the system to conduct its operations. It builds in checks all along the way. Path setup gets verified. Directory checks happen too. Images get validated before anything else. The model loads properly. Detection follows through. All this leads to a steady pipeline that users can rely on. Every choice point in the flowchart helps cut down on errors during runs. It boosts how tough the Multilingual Semantic Object Identification Using Computer Vision and NLP can be overall.

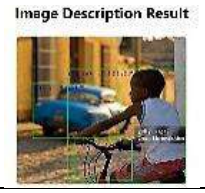
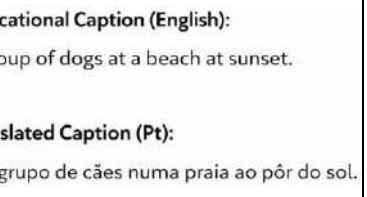
V. RESULTS AND DISCUSSION

A. System Testing

System testing really matters when it comes to checking the reliability, accuracy, performance, and usability of the Multilingual Semantic Object Identification Using Computer Vision and NLP in a thorough way. The main aim in this stage is to ensure that every component fits together seamlessly into a unified system. That covers things like object detection, language translation, speech synthesis, database handling, and user interface elements. All of them have to manage different kinds of situations without any problems. This testing checks that each part functions properly by itself. It also makes sure they interact effectively in actual real-time environments.

The setup for testing uses static images along with live camera feeds. It involves objects at various angles, distances, and spots to test detection reliability. Different lighting comes into play too, like natural daylight or dim indoor lights. Shadows and glare get checked to prove the system's strength. On top of that, the tests cover multiple languages, accents, and regional ways of speaking. All this helps measure translation precision and how clear the speech sounds.

Table 2: Unit test cases

Test Case Number	Input	Stage	Expected behavior	Observed behavior	Status P=Pass F=Fail
1	 Input: Objects Type of object	Input page	Object detected		P
2	Input: Objects Type of Language 	Input Page	Language Translated		P
3	Input: Objects description 	Input Page	Descriptions		P

B. Result Analysis

The main aim of the project was to detect the objects using machine learning algorithms. Table 3 shows the analysis that was performed on the three models with the different training and testing sizes. It was found that ensemble model was the most accurate in all the cases.

Table 3: Analysis of the three algorithms

TrainingSize	Testing Size	Accuracy		
		YOLOV3	Blip	Google Translate
70%	30%	82%	83%	80%
80%	20%	85%	87%	82%

In the below Figure 3 it shows the bar graph for the accuracy of the three algorithms where the trainset size was 80% and the test set size was 20%

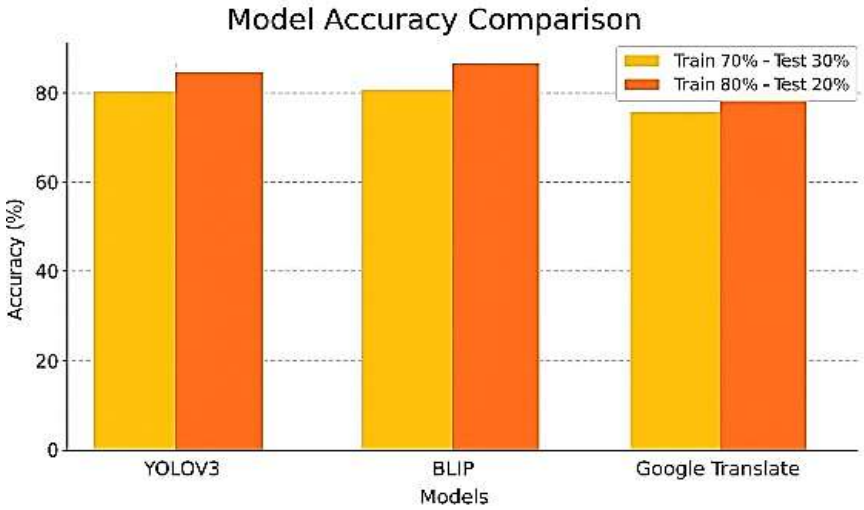


Figure 3: Graph analysis of the first set

VI. CONCLUSION

The study managed to show how they put together a system that picks out objects and makes sense of them in all sorts of languages. That tackles a big issue in places where people speak different tongues. They mixed in some cutting-edge ways to recognize images with tools for handling natural language. The setup turns out pretty adaptable. It spots objects and gives labels right in the local languages. Results point to how pairing deep learning for detecting objects with translation tweaks for each language boosts accuracy. It also makes things easier for users. In the end, this helps close the gap in talking across different language groups.

The framework has a modular architecture that really supports scalability. The approach allows the developers to add new languages and object categories easily without taking any major overhaul. Experiments demonstrate that the method holds steady performance on several datasets, proving its strength in practical settings. The paper points out how these systems could perform very well in education, assistance roles, and industry, especially in those areas where handling multiple languages is crucial.

Follow-up research will be able to further increase the vocabulary in the future. They might also push for a change in the pace of the processing times. Adding more context awareness would raise the accuracy regarding meanings. Loops added for user feedback might sharpen up the predictions and translations quite a bit. This effort, in the end, provides a good starting point for tools handling object recognition across multiple languages. The tool does deliver real practical value. It opens up space for further advances in AI systems that bridge languages.

VII. FUTURE WORK

The current system provides a strong foundation for recognizing objects across multiple languages, but there are several avenues for enhancement. The approach allows the developers to add new languages and object categories easily without taking any major overhaul. Experiments demonstrate that the method holds steady performance on several datasets, proving its strength in practical settings. The paper points out how these systems could perform very well in education, assistance roles, and industry, especially in those areas where handling multiple languages is crucial.

Follow-up research will be able to further increase the vocabulary in the future. They might also push for a change in the pace of the processing times. Adding more context awareness would raise the accuracy regarding meanings. Loops added for user feedback might sharpen up the predictions and translations quite a bit. This effort, in the end, provides a good starting point for tools handling object recognition across multiple languages. The tool does deliver real practical value. It opens up space for further advances in AI systems that bridge languages.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] William Perrizo, W. Jockheck, A. Perrera, D. Ren, and Y. Zhang, "Object boundary detection for ontology-based image classification," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 2, pp. 45–52, 2002. Available from: https://tinyurl.com/mr2fs35z

[2] Andreas Opelt, Axel Pinz, and Andrew Zisserman, "A boundary-fragment model for object detection," in *Computer Vision – ECCV 2006*, Lecture Notes in Computer Science, vol. 3954, Springer, Berlin, Heidelberg, 2006, pp. 575–588. Available from: https://link.springer.com/chapter/10.1007/11744047_44

[3] Muhammad Yasir, Md. Sakaouth Hossain, and Sulaiman Khan, "Object identification using manipulated edge detection techniques," *Science*, vol. 3, no. 1, pp. 1–7, 2022. Available from: https://tinyurl.com/zrc6ksmr

[4] R. Hussin, M. R. Juhar, W. Kang, and A. Kamarudin, "Digital image processing techniques for object detection from complex background image," *Procedia Engineering*, vol. 41, no. 4, pp. 340–346, 2012. Available from: https://doi.org/10.1016/j.proeng.2012.07.182

[5] P. Papageorgiou, M. Oren, and T. Poggio, "A general

framework for object detection," in *Proceedings of the Sixth International Conference on Computer Vision (ICCV)*,
[6] Bombay, India, 1998, pp. 555–562. Available from: https://tinyurl.com/3smzz4p5
[7] P. Viola and M. Jones, "Object detection using image processing," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 3, pp. 2875–2883, 2022. Available from: https://tinyurl.com/4mnmdh24
[8] S. Sankar Ganesh, K. Mohana Prasad, and Y. Karuna, "Object identification using wavelet transform," *Indian Journal of Science and Technology*, vol. 9, no. 5, pp. 1–6, 2016. Available from: https://tinyurl.com/42pamxha
[9] Qiankun Liu, Yichen Li, and Yuqi Jiang, "Generic multi-object tracking," *IEEE Transactions on Image Processing*, vol. 33, pp. 2456–2469, 2024. Available from: https://ieeexplore.ieee.org/abstract/document/10571840
[10] Samuele Buro and Isabella Mastroeni, "The multi-language construction," in *Formal Methods for Industrial Critical Systems*, Lecture Notes in Computer Science, vol. 12709, Springer, Cham, 2021, pp. 247–262. Available from: https://link.springer.com/chapter/10.1007/978-3-030-65474-0_14
[11] Zhong-Qiu Zhao, Shou-Tao Xu, and Xindong Wu, "Object detection with deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019. Available from: https://ieeexplore.ieee.org/abstract/document/8627998

ABOUT THE AUTHORS

Dr. Suchetha N V has completed her B.E in the year 2011, MTech in the year 2015 from VTU Belagavi, and Ph.D. in the year 2024 from VTU Belagavi. Her areas of interest are Image Processing and Deep Learning. She has published 13 international research papers. She is the life member of Kannada Sahitya Parishad.

Abhijeet P M received his B.E. degree in Computer Science and Engineering from Visvesvaraya Technological University (VTU), Belagavi. His areas of interest include Natural Language Processing (NLP) with Deep Learning and Artificial Intelligence (AI). He was a student in the Department of Computer Science and Engineering at SDM Institute of Technology, Ujire.

Likesh K received his B.E. degree in Computer Science and Engineering from Visvesvaraya Technological University (VTU), Belagavi. His area of interest includes Natural Language Processing (NLP) with Deep Learning and Artificial Intelligence (AI). He was a student in the Department of Computer Science and Engineering at SDM Institute of Technology, Ujire.

Vinuth A N received his B.E. degree in Computer Science and Engineering from Visvesvaraya Technological University (VTU), Belagavi. His area of interest includes Natural Language Processing (NLP) with Deep Learning and Artificial Intelligence (AI). He was a student in the Department of Computer Science and Engineering at SDM Institute of Technology, Ujire.

Suhas H received his B.E. degree in Computer Science and Engineering from Visvesvaraya Technological University (VTU), Belagavi. His area of interest includes Natural Language Processing (NLP) with Deep Learning and Artificial Intelligence (AI). He was a student in the Department of Computer Science and Engineering at SDM Institute of Technology, Ujire.