iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** XI **Month of publication:** November 2026

DOI: https://doi.org/10.22214/ijraset.2026.85686

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue XI Nov 2026- Available at www.ijraset.com

Pixenta: AI-Powered Intent-Aware Screenshot Assistant for Transforming Screenshots into Intelligent Actions

Sayyad Isma[1], Shivendra Pratap Singh[2], Siddhant Shukla[3]
Department of Computer Science and Engineering, Babu Banarasi Das Northern India Institute of Technology, Lucknow, Uttar Pradesh, India

Abstract: ***The widespread use of digital devices has made screenshots a convenient way to preserve information from shopping platforms, travel services, educational portals, event listings, and online communications. However, conventional gallery-based storage generally treats screenshots as images rather than as information-bearing records connected to a user's possible goals. This paper presents Pixenta, a prototype concept for an intent-aware screenshot assistant designed to move beyond passive image storage toward information extraction and contextual action recommendation. The current prototype provides a Streamlit-based interface for screenshot upload and uses Pillow for image handling and PaddleOCR for text extraction. A preliminary functional check was performed using a shopping screenshot, from which visible product and offer information was extracted. The broader proposed pipeline includes visual-context analysis, structured information extraction, semantic representation, similarity-based grouping, intent classification, and user-approved action recommendations such as price tracking or reminders. These advanced stages are presented as the system's proposed design and are not claimed as experimentally validated in the current prototype. The paper documents the motivation, architecture, preliminary implementation, evaluation plan, limitations, and future development of Pixenta.***

Keywords: Screenshot Understanding, Optical Character Recognition, Intent-Aware Computing, Semantic Organization, Action Recommendation.

I. INTRODUCTION

Screenshots are frequently used to retain information that appears temporarily on websites, mobile applications, and digital services. A user may capture a product listing to compare prices later, save a ticket to remember a journey, preserve an assignment notice to track a deadline, or capture an event announcement for future reference. As the number of screenshots grows, the gallery becomes a collection of visually stored but weakly structured information. The main limitation of conventional screenshot storage is that it generally supports storage and manual retrieval rather than understanding the information represented in an image. A user must often reopen an image, interpret its contents, remember why it was captured, and decide what action should follow. This creates friction between information capture and subsequent task completion. Pixenta is proposed as an AI-assisted screenshot management system that analyses screenshot content and, in its intended full workflow, infers a likely user intent before recommending a relevant action. The system is designed around a human-in-the-loop principle: recommendations should be presented for user review, and consequential actions should not be performed without user approval.

A. Problem Statement

The problem addressed in this work is the limited ability of ordinary screenshot collections to transform image-based information into structured, searchable, and actionable records. The research question is: how can text extraction and contextual analysis be combined in a screenshot-management workflow to support the identification of likely user intentions and the recommendation of relevant follow-up actions?

B. Objectives

- Develop a usable interface for uploading and processing screenshots.
- Extract text embedded in screenshots using optical character recognition.
- Represent extracted information in a structured form suitable for downstream processing.
- Design a pipeline for semantic comparison and grouping of related screenshots.

©IJRASET: All Rights are Reserved

- Design intent categories and contextual action recommendations for common screenshot scenarios.
- Evaluate the implemented components using a reproducible, labelled test set in future experimental work.

C. Scope of the Current Paper

This manuscript distinguishes between the implemented prototype and the proposed full system. The currently verified implementation consists of a Streamlit application, Pillow-based image handling, and PaddleOCR-based text extraction. A preliminary shopping-screenshot test is documented. Visual-language-model integration, embedding-based semantic grouping, automated intent classification, reminder scheduling, and notification delivery are described as proposed or future modules unless separately implemented and tested by the authors.

II. RELATED WORK

OCR converts text in screenshots into machine-readable information, providing Pixenta with textual content. However, it cannot independently identify user intent. Sentence-BERT enables semantic representation through sentence embeddings and similarity comparison, offering a potential approach for understanding screenshot context. Similarly, CLIP demonstrates how visual and textual information can be connected, which could improve the interpretation of images, icons, and layouts. DBSCAN offers a possible method for grouping similar screenshot embeddings and identifying outliers. These techniques provide a foundation for future Pixenta enhancements, but Sentence-BERT, CLIP, and DBSCAN remain proposed extensions rather than implemented features.

A. Research Motivation and Design Gap

The design motivation for Pixenta is to connect screenshot text extraction with later stages of semantic organization and task-oriented assistance. The contribution claimed in this paper is the design and preliminary implementation of a screenshot-processing prototype, together with a clearly specified extension architecture. A broad claim that no existing system performs similar functions is not made; such a claim would require a systematic and up-to-date review of commercial tools and academic literature.

III. PROPOSED SYSTEM AND METHODOLOGY

Pixenta is organized as a staged processing pipeline. Each stage produces an output that can be inspected or used by the next stage. The architecture is designed to support incremental implementation, allowing the OCR prototype to be extended without treating proposed modules as already operational.

1. Screenshot Input User uploads one or more screenshot images.
2. Image Handling Pillow loads and validates the image for processing.
3. OCR Extraction — IMPLEMENTED PaddleOCR extracts visible text from the screenshot.
4. Structured Information Extraction — PROPOSED Extract entities such as product names, prices, dates, locations, and deadlines.
5. Visual Context Analysis — PROPOSED A vision-language model may analyse visual elements and their relationship to extracted text.
6. Semantic Representation — PROPOSED A sentence-embedding model may convert extracted content into vector representations.
7. Similarity and Grouping — PROPOSED Similarity measures and a selected clustering method may group related screenshots.
8. Intent Classification — PROPOSED A classifier or rule-assisted model may estimate categories such as shopping, travel, study, and events.
9. Action Recommendation — PROPOSED The system may suggest actions such as price tracking or deadline reminders.
10. User Review The user reviews, edits, accepts, or rejects a recommendation.
11. Reminder / Follow-up — FUTURE WORK Approved reminders or integrations may be scheduled and delivered.

Figure 1. Pixenta's staged system workflow. Green shading identifies the currently verified OCR implementation; yellow shading identifies proposed or future modules.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

A. *Screenshot Acquisition and Image Handling*

The user interface is implemented using Streamlit. Uploaded image files are handled using the Python imaging library Pillow. The input stage is responsible for receiving screenshot files and making them available to the OCR component. For a production implementation, file type validation, file-size limits, corrupted-image handling, and safe temporary-file management should be enforced.

B. *Optical Character Recognition*

PaddleOCR is used to identify and recognize text present in an uploaded screenshot. The output is a collection of recognized text segments that can be displayed to the user or passed to later processing stages. OCR quality may be affected by low resolution, compression, skew, small fonts, unusual typography, overlapping elements, and low contrast. The current prototype demonstrates text extraction but has not yet been evaluated on a sufficiently large annotated dataset.

C. *Structured Information Extraction — Proposed*

A future information-extraction module may identify entities and attributes such as product title, brand, listed price, discount, travel date, departure time, venue, or assignment deadline. Extracted values should retain their source text and, where possible, confidence information so that users can verify them before acting.

D. *Visual Context Analysis — Proposed*

OCR may not capture all relevant meaning in a screenshot. A future vision-language model could provide a visual description or classify the broad type of screen. The visual output should be combined with OCR text rather than treated as unquestionable ground truth. The model, prompt, inference settings, and error-handling strategy must be documented when this component is implemented.

E. *Semantic Representation and Similarity — Proposed*

After cleaning and structuring extracted text, a sentence-embedding model may map each screenshot's textual content to a vector. The cosine similarity between two vectors A and B can be expressed as:

$$\text{cosine_similarity}(A, B) = (A \cdot B) / (\|A\| \times \|B\|)$$

The score can be used to rank semantically related screenshots. The embedding model, text normalization, similarity threshold, and treatment of short or empty OCR outputs must be selected and evaluated empirically.

F. *Clustering and Intent Classification — Proposed*

A clustering algorithm may group screenshots whose extracted content is semantically related. DBSCAN is one candidate because it can identify dense groups and mark outliers as noise . Intent classification may then assign a likely category, for example shopping, travel, education, or events.

A hybrid approach combining rules and a trained classifier may be considered, but its design should be justified by measured performance.

G. *Action Recommendation and Human Approval — Proposed*

The recommendation stage maps an estimated intent and extracted entities to a possible action. For example, a product screenshot may lead to a price-tracking suggestion, while a ticket screenshot may lead to a journey reminder. The recommendation must be presented as a suggestion, not as certainty. The user should be able to approve, edit, dismiss, or delete it. Any external monitoring, notification, or account integration requires explicit user authorization and appropriate security controls.

IV. PROTOTYPE IMPLEMENTATION

The verified prototype uses Python and Streamlit to provide an interactive interface for screenshot upload and OCR processing. Pillow is used for image handling, while PaddleOCR performs text recognition. The implementation demonstrates the initial feasibility of extracting machine-readable text from a real-world shopping screenshot.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

A. *Technology Stack*

Component	Technology	Role / Status
Application interface	Streamlit	Implemented
Image handling	Pillow	Implemented
Text extraction	PaddleOCR	Implemented
Programming language	Python	Used by prototype
Visual-language model	Not confirmed	Proposed
Sentence embeddings	Not confirmed	Proposed
Clustering	Not confirmed	Proposed
Reminder / notification engine	Not confirmed	Future work

B. *Preliminary Functional Check*

A shopping screenshot from the Flipkart platform was used for a preliminary OCR demonstration. The previously recorded output included the platform name and visible product attributes, including the product text 'CRAETA sneakers', colour 'black', size '7', a listed price of ₹1,499, another displayed amount of ₹425, a displayed discount of 72%, and an offer amount of ₹342. These values are reported as the observed OCR demonstration output, not as independently verified commercial facts. The original screenshot and raw OCR output should be retained as supporting evidence.

This single demonstration confirms that the prototype can extract selected visible text from at least one tested screenshot. It does not establish general OCR accuracy, robust product extraction, intent-classification accuracy, clustering quality, or reminder reliability.

V. EVALUATION PLAN

A rigorous evaluation is required before making quantitative performance claims. The following plan is proposed for the next development stage.

Module	Suggested measure	Required evidence
OCR	Character Error Rate (CER), Word Error Rate (WER)	Ground-truth transcription for each test image
Intent classification	Accuracy, precision, recall, F1-score	Manually labelled intent classes and predictions
Clustering	Silhouette score, cluster purity where appropriate	Embeddings, assigned clusters, reference labels
Recommendations	Precision@K or human relevance rating	Judgements of whether suggested actions are useful
Runtime	Mean and median processing time	Repeated timed runs on documented hardware
Usability	Task completion and user feedback	Defined user test protocol and consent

The dataset should contain diverse screenshot types and should be annotated by humans using a written labelling guide. The paper should report the number of images, category distribution, inclusion and exclusion criteria, train/validation/test split where model training is performed, hardware, software versions, and evaluation procedure. Personal information should be removed or anonymized. No performance values are inserted in this draft because a verified experimental dataset and measured results have not been provided.

A. *Baseline and Ablation Comparisons*

Useful comparisons include OCR text with rule-based categorization, OCR combined with semantic embeddings, and a multimodal approach combining OCR with visual analysis. An ablation study may remove one component at a time to measure its contribution. These comparisons should be performed only after the corresponding modules are implemented.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VI. DISCUSSION

The preliminary prototype demonstrates the first stage of the intended workflow: converting screenshot text into machine-readable output through OCR. This is a necessary foundation for later information extraction and semantic processing. However, text extraction should not be confused with understanding a user's intention. A screenshot may contain several possible purposes, and the same visible content can lead to different actions depending on the user's context.

The proposed human-in-the-loop design is important because incorrect extraction or intent prediction could create misleading reminders or actions. Users should be shown the extracted values and the reason for each recommendation. Confidence thresholds, confirmation steps, and the ability to correct errors should be considered during implementation.

The current evidence is preliminary and limited to a demonstrated shopping screenshot. Therefore, the paper does not claim that the complete intent-aware pipeline has been validated or that Pixenta outperforms existing screenshot-management tools. Those conclusions require a larger comparative evaluation.

VII. LIMITATIONS AND FUTURE WORK

1) The current verified prototype is limited to screenshot upload, image handling, and OCR text extraction.
2) No large, manually annotated screenshot dataset or quantitative benchmark has yet been documented.
3) Visual understanding, semantic embeddings, clustering, automated intent classification, and reminder delivery require implementation and testing.
4) OCR may fail on blurred, low-resolution, multilingual, or visually complex screenshots.
5) Intent can be ambiguous and may depend on personal context unavailable from the screenshot itself.
6) Future work includes privacy-preserving storage, multilingual processing, confidence-aware recommendations, user studies, and integrations with calendar or price-monitoring services.

VIII. CONCLUSION

This paper has presented Pixenta, an intent-aware screenshot assistant proposed to transform passive screenshot collections into structured information and context-sensitive action suggestions. The current prototype establishes an initial implementation using Streamlit, Pillow, and PaddleOCR, and a preliminary shopping-screenshot demonstration illustrates text extraction from a real-world image. The broader architecture proposes visual-context analysis, structured information extraction, semantic representation, screenshot grouping, intent classification, and user-approved recommendations. These stages provide a development roadmap rather than verified experimental outcomes. Future work will implement and evaluate the complete pipeline using a labelled dataset, quantitative metrics, baseline comparisons, and user feedback. Such evaluation is necessary to determine the system's reliability and practical value.

IX. ACKNOWLEDGMENT

The authors express their sincere gratitude to their project guide, Mr. K. S. Pathak, for guidance and academic support during the development of the Pixenta project. The authors also acknowledge the support of their department and institution.

A. *Author Contributions*

The author list is recorded as provided by the project team. The specific contribution statement has not been assigned in this draft because individual task records were not available. Before submission, the authors should agree on and insert an accurate contribution statement covering conceptualization, software development, investigation, validation, writing, and supervision. Authorship should reflect actual scholarly contributions and institutional policy.

B. *Declarations*

Funding: This research received no external funding. The development and preparation of the Pixenta prototype were carried out independently for academic and research purposes.

Conflict of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work presented in this paper.

Data availability: The data used for the development and evaluation of the Pixenta prototype consist of screenshot samples and OCR-generated text records. Anonymized sample data and OCR evaluation records may be made available by the corresponding author upon reasonable request, subject to privacy and confidentiality considerations. Original screenshots containing personal, sensitive, or identifiable information are not publicly shared to protect user privacy.

©IJRASET: All Rights are Reserved

Ethics and privacy: Pixenta prioritizes user privacy by ensuring screenshots are processed only for their intended purpose. Personal and sensitive information should be handled securely, with user consent and without unauthorized sharing.

Use of AI tools: AI tools were used for language refinement, grammar correction, and improving the clarity of this research paper. The authors are responsible for verifying the content and ensuring its accuracy and originality.

REFERENCES

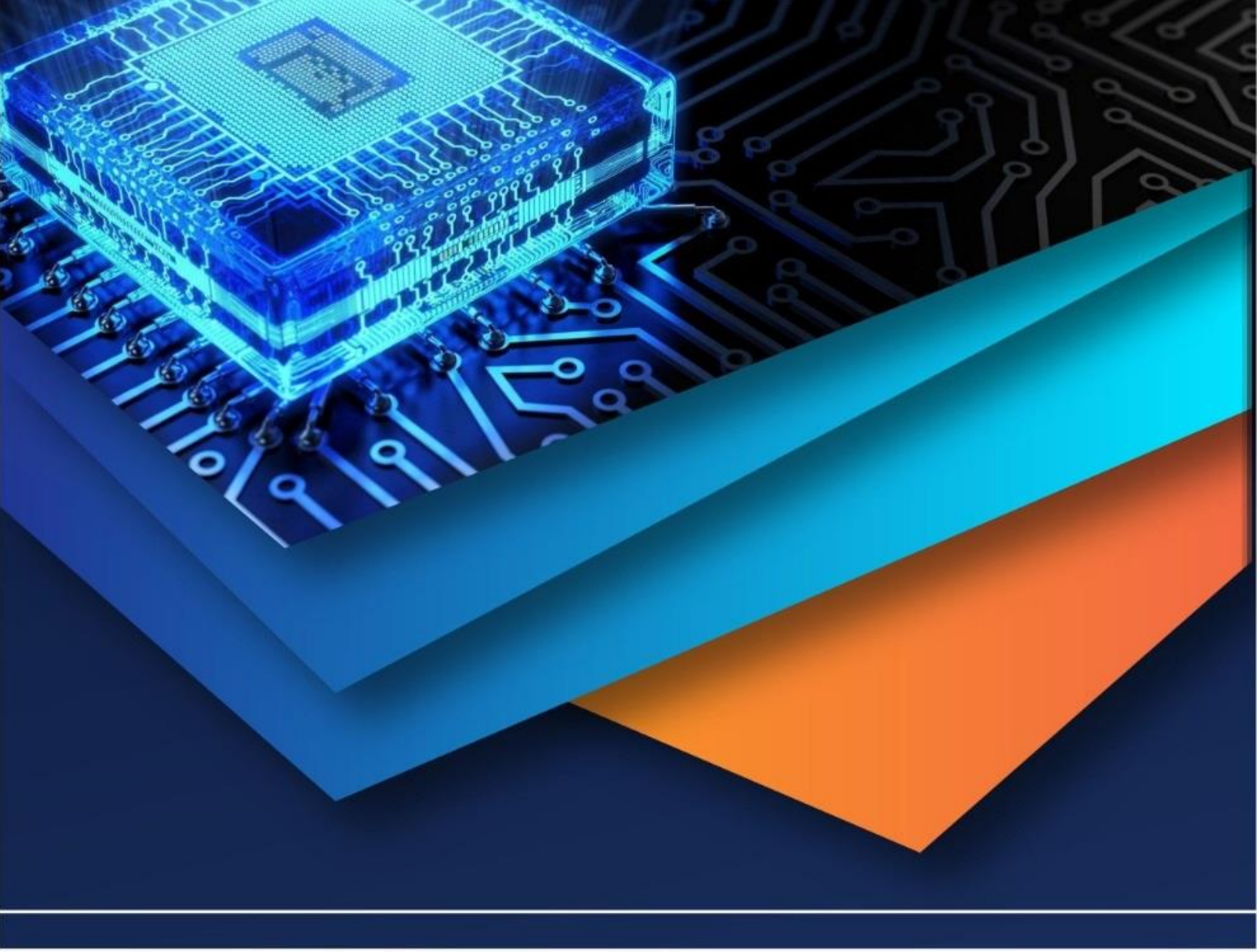
[1] Y. Du, C. Li, R. Guo, X. Yin, W. Liu, J. Zhou, Y. Bai, Z. Yu, Y. Yang, Q. Dang, and H. Wang, "PP-OCR: A Practical Ultra Lightweight OCR System," arXiv:2009.09941, 2020. Available: https://arxiv.org/abs/2009.09941

[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

[3] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proceedings of the 38th International Conference on Machine Learning, PMLR, vol. 139, 2021, pp. 8748–8763. Available: https://proceedings.mlr.press/v139/radford21a.html

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996, pp. 226–231.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)