ImageSeek: A Hybrid Text-to-Image Image Retrieval System for Domain-Specific Collections

Rodrigo Duarte[1,2(✉)], Rodrigo Silva[1,2], António Branco[4,5], Hugo Proença[1,6], and Ricardo Campos[1,2,3]

[1] University of Beira Interior, Covilhã, Portugal
{rodrigo.duarte,rd.silva,hugomcp,ricardo.campos}@ubi.pt
[2] INESC TEC, Porto, Portugal
[3] Ci2 - Smart Cities Research Center, Polytechnic Institute of Tomar, Tomar, Portugal
[4] University of Lisbon, Lisbon, Portugal
antonio.branco@di.fc.ul.pt
[5] NLX-Group, Lisbon, Portugal
[6] IT: Instituto de Telecomunicações, Covilhã, Portugal

Abstract. Large image collections are typically organized around basic metadata and keyword tags, making content discovery challenging for users seeking specific visual information. Although images may be accompanied by descriptive text, traditional retrieval systems often struggle to bridge the semantic gap between textual descriptions and visual content. In this demo, we present ImageSeek, a hybrid text-to-image retrieval system designed to enhance search effectiveness by combining text and image-based retrieval methods through an asymmetric score adjustment mechanism. The system leverages multilingual CLIP models to encode both visual and textual information, creating unified representations for cross-modal retrieval. Users can search through natural language queries in any supported language, with results ranked using a hybrid approach that treats image-based retrieval as a reliable baseline while harmonizing text-based scores through position-dependent adjustments. The demonstration system operates on a dataset of 42,333 images from the Portuguese Presidency website, providing an appropriate testbed for multimodal retrieval performance. The web application enables direct comparison between conventional CLIP-based retrieval and our hybrid approach, supporting image searches under the same conditions on external platforms, including Google Images and the Arquivo.pt image search system, enabling comparative analysis of the results. To evaluate its effectiveness, ImageSeek allows users to experience differences between retrieval modes while exploring domain-specific visual content.

Keywords: Image Information Retrieval · Image Search · CLIP

R. Campos et al. (Eds.): ECIR 2026, LNCS 16486, pp. 166–171, 2026.
https://doi.org/10.1007/978-3-032-21321-1_24

1 Introduction

The rapid growth of visual content across digital platforms has created significant challenges in effectively retrieving relevant images from large collections [1]. Cross-modal information retrieval (IR) systems have emerged to address these challenges [2], with text-to-image IR systems gaining significant attention through vision-language models like CLIP [3]. Recent advances have explored entity-aware approaches that inject semantic information into multimodal representations [4], demonstrating the potential for domain adaptation of foundation models. However, most existing systems focus on English-language queries and general-purpose image collections, leaving a gap for specialized domains and multilingual contexts. This limitation is particularly problematic for non-English languages, where translation-based strategies often fail to preserve nuanced meanings and cultural context [5]. In this paper, we present a demonstration of a hybrid text-to-image retrieval system that integrates text and image-based retrieval methods. While our demonstration uses Portuguese presidential archives as a testbed, the system is designed to be domain-agnostic and can be applied to any multimodal collection.

The key innovation of our system lies in its asymmetric score adjustment mechanism. Unlike traditional rank fusion approaches [6] our method treats image-based retrieval as a reliable baseline while harmonizing text-based scores through position-dependent adjustments, enhancing contextual relevance through textual information. Our demonstration operates on a comprehensive dataset [7][1] of 42,333 images collected from the Portuguese Presidency website[2]. The dataset includes 80 Portuguese textual queries, each associated with relevance judgments for an average of 65 images. Annotation was conducted by three master's students in computer science, who independently assigned binary relevance labels (relevant/non-relevant) to each query-image pair. Inter-annotator agreement was measured using Fleiss' Kappa (K=0.62), indicating moderate agreement. Conflicting annotations were resolved through majority voting. In total, 5,201 images were manually annotated for relevance, providing a robust testbed for evaluating text-to-image retrieval performance in domain-specific scenarios. The web application enables users to experience the differences between conventional CLIP-based retrieval and our hybrid approach, alongside with searches on Google Images and Arquivo.pt [8]. ImageSeek is publicly available online[3] with the full source code on GitHub[4], supporting reproducibility and enabling further research development. A demonstration video of the platform is available on the demo website.

[1] https://github.com/LIAAD/pt-image-ir-dataset.
[2] https://www.presidencia.pt/.
[3] https://imageseek.inesctec.pt/.
[4] https://github.com/LIAAD/imageseek-demo.

2 Architecture

Our system is implemented using Flask, with Redis Stack serving as a key-value store and vector database for efficient storage and retrieval of image embeddings, article embeddings, and metadata. For textual representation, the system indexes images using article titles as metadata. While more comprehensive textual representations could provide richer semantic information, article titles offer several practical advantages: they are consistently available across the collection, provide high-level topical context, and enable efficient indexing and retrieval. Pre-computed embeddings for all 42,333 images and their corresponding article titles are generated using multilingual CLIP models [5]. Specifically, the system uses the OpenCLIP xlm-roberta-base-ViT-B-32 model trained on LAION-5B [9] to generate both text and image embeddings, ensuring consistent cross-modal representations within the same vector space. This model supports cross-lingual retrieval for Portuguese queries while preserving semantic alignment for the hybrid scoring mechanism. While our demonstration focuses on Portuguese presidential archives, the system's architecture is inherently language-agnostic and domain-independent. The asymmetric score adjustment mechanism operates on normalized similarity scores and relative positions, making no language-specific or domain-specific assumptions. The underlying multilingual CLIP model supports over 100 languages through its XLM-RoBERTa text encoder, enabling zero-shot transfer to other languages without modification. The three-phase algorithm depends only on textual metadata associated with images, a vision-language model producing comparable embeddings, and score distributions that can benefit from harmonization—conditions satisfied in most multimodal collections. The core innovation is a three-phase hybrid retrieval algorithm (see Fig. 1) that addresses the key challenge of multimodal score harmonization.

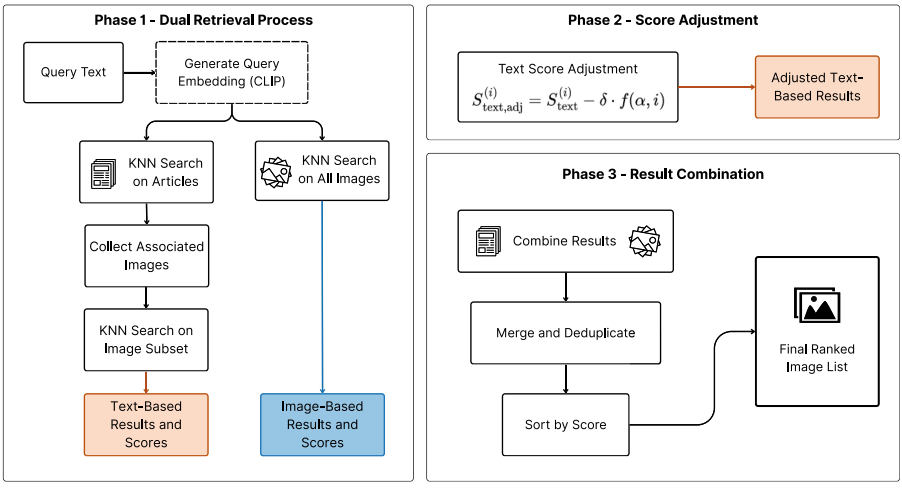

Fig. 1. Three-phase hybrid retrieval algorithm of ImageSeek.

Phase 1 begins by generating a text embedding for the input query using the selected multilingual CLIP model. Dual retrieval is then performed in parallel using K-Nearest Neighbors (KNN) searches. The text-based approach first retrieves semantically similar articles by comparing the query embedding against article title embeddings, and then searches for images within them. Simultaneously, the image-based approach performs KNN directly against all precomputed image embeddings. Both produce ranked lists with cosine distance scores on different search spaces, yielding distinct score distributions. **Phase 2** applies asymmetric score adjustment to harmonize the modality-specific scores, treating image-based scores as a reliable baseline, and adjusting only text-based scores using position-dependent functions: $S^{(i)}_{text,adj} = S^{(i)}_{text} - \delta \cdot f(\alpha, i)$ where $\delta = S^{(1)}_{text} - S^{(1)}_{image}$ represents the systematic offset between modalities, addressing the fundamental challenge that text-based and image-based retrieval produce scores on different scales and distributions. The δ term captures the systematic offset between top-ranked items from each modality, while $f(\alpha, i)$ applies position-dependent adjustment that gradually diminishes the influence of textual signals at lower ranks, reflecting the hypothesis that textual context is most valuable for refining top-ranked visual matches, enabling effective multimodal score fusion. Four adjustment function variants are implemented: linear with zero indexing ($f(\alpha, i) = 1 - -\alpha \cdot i$), linear with one indexing ($f(\alpha, i) = 1 - \alpha \cdot (i-1)$), square root adjustment ($f(\alpha, i) = 1 - \alpha^{\sqrt{i-1}}$), and exponential adjustment ($f(\alpha, i) = 1 - \alpha^{e^{i-1}}$), allowing users to explore different score harmonization strategies. To compare these variants, we evaluated their performance across adjustment factor values ranging from $\alpha = 0.0$ (no adjustment) to α=1.0 (maximum adjustment). Linear zero-indexed consistently achieved the best results, with MRR = 0.621 at $\alpha = 0.1$ and F1@10 = 0.294 at $\alpha = 0.2$. This function demonstrated stable performance across the $\alpha = 0.1$–0.3 range, with minimal variation (MRR variance < 0.005). In contrast, non-linear functions showed higher sensitivity to parameter changes, with square root performance declining from MRR = 0.619 to 0.610 as α increased from 0.0 to 1.0, and exponential adjustment exhibiting similar degradation patterns. Linear one-indexed produced comparable but slightly lower results (MRR = 0.617, F1@10 = 0.293 at $\alpha = 0.1$). The superior stability of linear zero-indexed, combined with its computational simplicity and interpretability, motivated its selection as the default configuration. **Phase 3** combines the adjusted text-based results with unchanged image-based results, removes duplicates by keeping the better score per image, and produces a final unified ranking sorted by distance scores.

3 Demonstration

The web interface provides an intuitive platform for exploring our hybrid retrieval approach on the previously described dataset. To demonstrate the system's effectiveness, we showcase a comparative analysis of our system and Google image results using the query "Presidente a ler um livro" (President reading a book). Figure 2 shows the search interface with mode selection options. For this demonstration, the hybrid Linear Decay (Zero-indexed) mode serves as the

default configuration, as preliminary evaluations indicate it yields the most effective results. Users can toggle between conventional CLIP-based retrieval, our hybrid approach, and external service comparisons, such as Google Images and Arquivo.pt (the Portuguese web archive). External comparisons are restricted to results from the Portuguese Presidency website domain and matched to the same temporal range to ensure fair comparison within a consistent content scope.

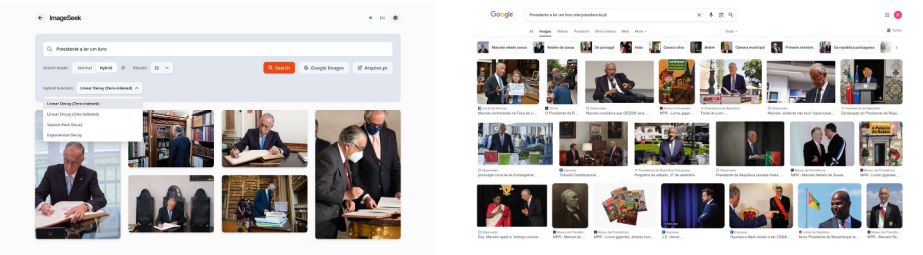

Fig. 2. Search interface showing hybrid mode results (left) and Google Images comparison for the same query (right).

The comparison reveals the system's effectiveness in retrieving contextually relevant images from the Portuguese Presidency domain. While Google Images returns broader, often less context-aware results, our hybrid approach leverages domain-specific knowledge and asymmetric score adjustment to produce more precise and semantically aligned matches. This demonstrates how domain-adapted hybrid retrieval can outperform general-purpose systems in specialized collections. Furthermore, the system offers researchers an interactive platform to experiment with different score adjustment strategies and explore trade-offs between text and image-based retrieval modalities in real-time, fostering further research in domain-specific text-to-image retrieval systems.

In addition to these comparative insights, we also present a quantitative evaluation. Table 1 reports evaluation results on the *pt-image-ir-dataset*, demonstrating that our hybrid approach consistently outperforms baseline methods across all metrics.

Table 1. Performance comparison of our hybrid approach and baseline methods on the pt-image-ir-dataset. Best results in each column are highlighted in bold.

Method	MAP	P@5	R@5	P@10	R@10	F1@10	MRR	RP
Google Images	0.076	0.202	0.076	0.190	0.125	0.128	0.336	0.113
Arquivo.pt	0.038	0.145	0.045	0.134	0.080	0.091	0.217	0.072
OpenCLIP xlm-roberta-base	0.176	0.418	0.121	0.419	0.264	0.288	0.610	0.218
Hybrid Approach	**0.179**	**0.420**	0.125	**0.423**	**0.268**	**0.293**	**0.621**	0.221
RRF Method	0.171	0.413	**0.127**	0.414	0.261	0.289	0.597	**0.222**

Acknowledgments. The authors Rodrigo Duarte, Rodrigo Silva and António Branco would like to acknowledge ACCELERAT.AI - Multilingual Intelligent Contact Centers, funded by the covid-recovery program PRR-Plano de Recuperação e Resiliência, through IAPMEI (C625734525-00462629); PORTULAN CLARIN - Research Infrastructure for the Science and Technology of Language, funded by LISBOA2030 (FEDER-01316900); hey, Hal, curb your hallucination!, funded by FCT-Fundação para a Ciência e Tecnologia (2024.07592.IACDC). The author Hugo Proença would like to acknowledge FCT – Fundação para a Ciência e a Tecnologia, I.P., and, when eligible, co-funded by EU funds under project/support UID/50008/2025 – Instituto de Telecomunicações, with DOI identifier <https://doi.org/10.54499/UID/50008/2025> Ricardo Campos is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (https://doi.org/10.54499/UID/50014/2025). The author would also like to acknowledge project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Mourão, A., Gomes, D.: Searching images in a web archive. In: 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10. IEEE (2023)
2. Gong, Y., Cosma, G., Finke, A.: VITR: augmenting vision transformers with relation-focused learning for cross-modal information retrieval. ACM Trans. Knowl. Discov. Data **18**(9), 1–21 (2024)
3. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748–8763. PMLR (2021)
4. Adjali, O., Ferret, O., Ghannay, S., Le Borgne, H.: Entity-aware cross-modal pretraining for knowledge-based visual question answering. In: Hauff, C., et al. (eds.) Advances in Information Retrieval, pp. 391–400. Springer Nature Switzerland, Cham (2025)
5. Carlsson, F., Eisen, P., Rekathati, F., Sahlgren, M.: Cross-lingual and multilingual clip. In: Proceedings of the thirteenth language resources and evaluation conference, pp. 6848–6854 (2022)
6. Cormack, G.V., Clarke, C.L.A., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 758–759 (2009)
7. Duarte, R., Branco, A., Proença, H., Campos, R.: pt-image-ir-dataset: An Image Retrieval Dataset in European Portuguese. In: Anand, A., et al. (eds.) Lecture Notes in Computer Science - Advances in Information Retrieval - ECIR'26 - 48th European Conference on Information Retrieval. Springer, Delft, Netherlands (2026)
8. Gomes, D., Cruz, D., Miranda, J., Costa, M., Fontes, S.: Search the past with the portuguese web archive. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 321–324. ACM, New York, NY, USA (2013)
9. Ilharco, G., et al.: OpenCLIP. Zenodo (2021). https://doi.org/10.5281/zenodo.5143773