H-EAGLE: Hierarchical Extension of EAGLE for Multi-Level Semantic Video Retrieval

Thang-Long Nguyen-Ho[1], Viet-Tham Huynh[2,3], Allie Tran[1], Minh-Triet Tran[2,3], Cathal Gurrin[1], and Graham Healy[1]

[1] Dublin City University, Dublin, Ireland
[2] Software Engineering Laboratory, University of Science, VNU-HCM, Vietnam
[3] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract. Modern Video Retrieval systems face challenges in computational efficiency and semantic depth when handling complex queries, particularly those with time-sensitive requirements. These systems typically rely on a "flat" index structure that encodes each frame independently, resulting in high search costs and difficulty capturing higher-level events or context semantics. To address these limitations, we propose a novel three-level hierarchical index concept that organizes video data at different semantic abstraction levels. The first level involves embedding vectors for individual frames to facilitate fine-grained retrieval. The second level groups visually similar frames into "shots" and encodes them into a semantic temporal representation. The top layer uses a Visual-Language Model (VLM) to identify and group frames related to narrative actions. This architecture allows the system to first quickly identify high-level related scenes or actions, and then refine the results by searching within individual frames within those groups. Our approach helps users to query data at the most relevant conceptual level.

Keywords: Video Retrieval · Temporal Retrieval · Retrieval System · Hierarchical Indexing

1 Introduction

The growth progress of Video Retrieval [21] over the past decade has demonstrated the representational power of foundation models, beginning with the image–text embedding model CLIP [14] and its subsequent variants. These models have successfully helped to bridge the semantic gap between text and images at the object level. However, this progress in feature representation has not been matched by progress in feature organization, as indexing architectures remain largely rudimentary. Although flat indices are computationally inexpensive, they are poorly suited to the complexity of real user queries, thereby limiting the practical capabilities of modern retrieval systems.

From an optimization perspective, finding a sequence of events that respects time constraints on a set of independent feature vectors is a hard problem. Heuristic methods that incorporate dynamic programming can yield feasible

solutions, but their computational costs increase exponentially with the size of the searching space. As a result, the performance of these systems is limited. From a representation architecture perspective, the flat structure fails to model the inherent hierarchical structure of real-world events. A query "birthday party" does not refer to the collection of discrete images, but a structure consisting of sub-events, such as "blowing candles" and "opening presents", each of which is composed of elementary images. Returning only isolated related frames not only degrades the user experience but also shows a weakness in understanding semantics at the macro level.

Recognizing these inherent challenges, we hypothesize that imposing a hierarchical structure on the feature space can address both computational efficiency and semantic depth. In this work, we present a three-level index concept that decomposes and represents data at different levels of abstraction, including instantaneous visual features, spatiotemporally coherent shots, and narrative actions identified by video understanding models..

Our system concept moves from "pattern matching", where other teams focus on exploiting the "pattern" from the query to the target images, to "contextual reasoning", where we not only consider the visual details but also the dependency contexts. Our main contributions include proposing a multi-level encoding strategy and retrieval algorithm, resulting in a scalable system.

2 Related Works

Interactive Video Retrieval (IVR) [21] has evolved from keyword-based search to multi-modal. The annual Video Browser Showdown (VBS) [15][16][17][18] serves as a benchmark for retrieval systems that integrate the latest techniques on video retrieval into complete systems. Some of the key research directions that shape the current landscape are presented in this section.

Representation Learning Integeration The advent of multimodal encoding models such as CLIP [14] and later encoders such as ALIGN [5], Florence [23], and BLIP [9] forms the backbone of most modern retrieval systems, providing the ability to embed text-images without processing, which is essential for semantic search. By considering videos as collections of frames, systems such as VideoEase [20] and VERGE [13] focus on merging and optimally weighting features from multiple embedding models to improve retrieval accuracy. This approach focuses on efficient use and combination of available representations.

Enhanced data usage using VLM/LLM This integration takes two main forms. First, LLM is used to bridge the semantic gap between user intent and query formulation. Leading systems such as NII-UIT [3], HORUS [10], and ViewsInsight2.0 [22] leverage models such as GPT-4 [12] to perform query expansion and rewrite. This process helps clarify user queries, handle ambiguity or misspellings, and enrich search results with relevant synonyms and contextual concepts, significantly improving the recall performance of the embedding model.

Second, generative models are used for data enrichment. The work Interactive Video Search with Multi-modal LLM Video Captioning [1] uses VLM to generate millions of descriptive captions for large-scale video datasets. These generatively constructed pairs are then used for the retrieval model, enhancing the capabilities of the original model.

Feature search with time constraint As retrieval tasks become more complex, the need for precision beyond the frame level increases. Temporal search has emerged as an important feature for tasks involving event sequences. Systems such as NII-UIT [3] introduce dynamic temporal search, which allows the user to define a series of actions or events, which the system then attempts to match in sequence within the video repository, avoiding the need to query single, isolated events.

Human Computer Interaction. In parallel with platform advances, significant innovations are taking place at the user interface level to create more natural and efficient search experiences. A radical departure from traditional 2D interfaces is VR work. LUMINA-1 [7] and vitrivr-VR [19], respectively, use VR/MR to display search results in interactive 3D spaces, allowing for intuitive exploration through gestures and eye movements. On conventional platforms, new interaction methods are also being introduced. VEAGLE [11] pioneered the use of eye tracking as a form of implicit relevance feedback, re-ranking results based on active attention. Meanwhile, systems such as Exquisitor [6] and SnapSeek 2.0[4] continue to refine conversational search loops and user relevance feedback, allowing for collaborative dialogue between users and systems to incrementally improve results. These advances emphasize the importance of quality and intuitiveness in user interaction design.

Overall, the direction of progress points to multi-step retrieval strategies that leverage temporal coherence and contextual information to progressively refine search targets.

3 System Overview

Current image retrieval systems are primarily built on a naive frame-level index, which suffers from several key gaps: they struggle to efficiently search over long temporal horizons, fail to capture semantic structure, and rely on search strategies that are often unintuitive.

To address these issues, we extend the naive index with two complementary layers representing shots and narrative actions. Shots are generated using traditional shot boundary detection methods to group visually coherent frames, while narrative actions are extracted via state-of-the-art video understanding models to segment semantically meaningful events. These two layers are complementary and non-overlapping, enabling the system to efficiently handle both low-level visual continuity and high-level event semantics. This hierarchical design improves retrieval efficiency, semantic understanding, and search flexibility. Figure 1 illustrates our data processing and query workflow.

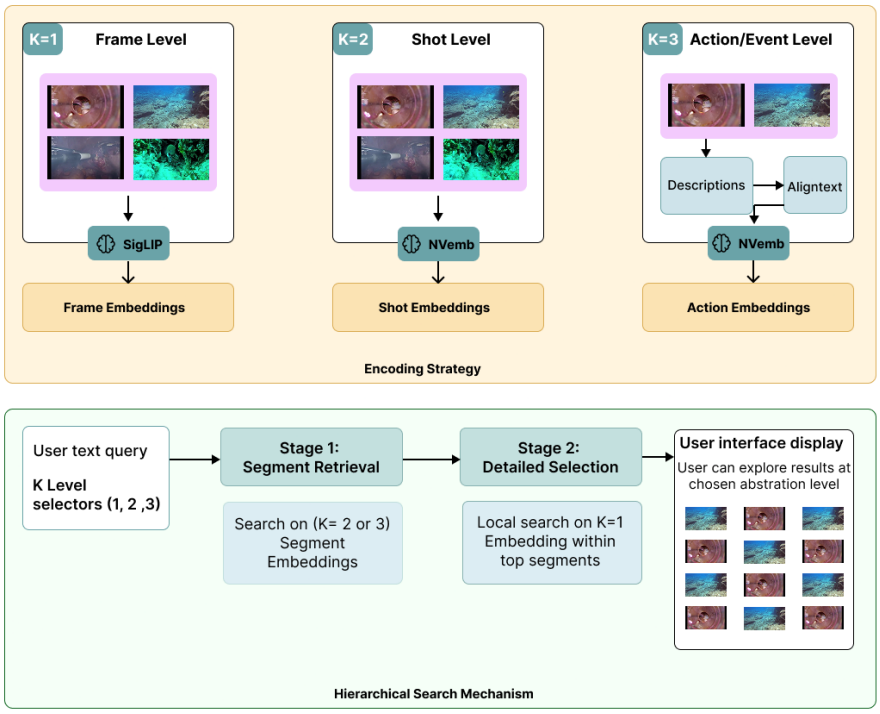

Fig. 1. The hierarchical architecture of our retrieval system. The dataset is indexed at three distinct levels: **Level 1** ($K = 1$) consists of individual image embeddings using SigCLIP. **Level 2** ($K = 2$) groups visually similar images into "shots", which are then embedded using a video encoder (NVemb). **Level 3** ($K = 3$) groups high-level "actions" or events extracted by VLM, which are embedded using NVemb. User queries can target any combination of these layers, allowing for flexible and efficient multi-stage searching.

3.1 Encoding Strategy

Our core contribution is a three-level embedding strategy that creates a rich, structured representation of the image dataset, as illustrated in Figure 1. Each level, or layer, corresponds to a different level of semantic granularity.

Level 1: Image-Level Representation ($K = 1$). This foundational layer provides the most detailed representation of the data. Each individual image in the dataset is processed by a powerful pre-trained visual language model, SigLIP [24], to produce an embedding vector. This embedding captures the semantic content and visible features of a single moment. This layer is essential for accurate instance-level retrieval, such as finding a specific object or instant action.

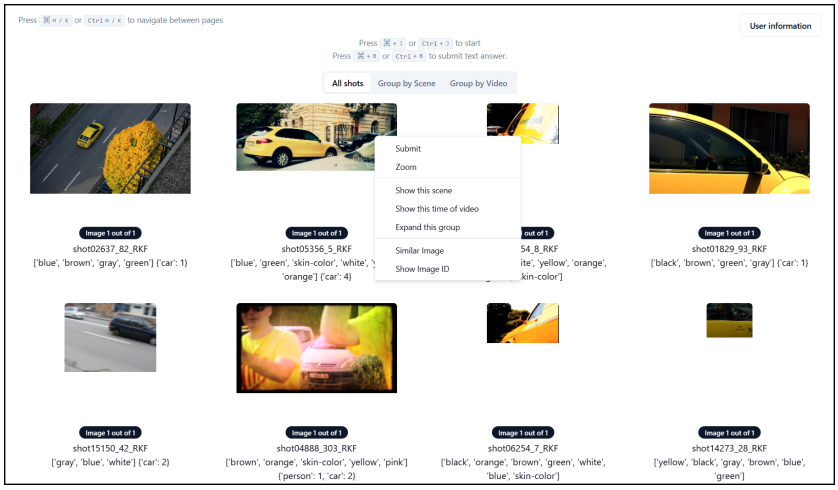

Fig. 2. A screenshot of the system user interface

Level 2: Shot-level abstraction ($K = 2$). Videos are typically characterized by sequences of similar images, called shots. To capture this structure, we build a shot-level representation, leveraging shot boundary detection techniques. We then form clusters by grouping these keyframes with images that are temporally and visually close to each other, subject to a high similarity threshold. Each resulting cluster, representing a coherent "shot" is considered a short sequence of images. This sequence is then fed into a video embedding model, NVemb [8], which has demonstrated its effectiveness on almost benchmarks.

Level 3: action level abstraction ($K = 3$). The generalization level in our hierarchy aims to group images based on common human activities or events, which often span multiple distinct visual shots (e.g., "baking a cake" might include scenes from the kitchen and dining room). To achieve this, we use a state-of-the-art Visual Language Model (VLM) to generate textual descriptions of potential activities occurring in the dataset. For each generated activity description, we use an advanced alignment method [2] to identify and match all corresponding images from the dataset. This semantically defined group of images, representing a complete action, is then encoded into a sequence using the same video embedding model, NVemb [8], to create a high-level event embedding.

3.2 Hierarchical Search Mechanism and Interface

This multi-level indexing structure enables an extremely flexible and efficient search mechanism. The user is empowered to specify the semantic level(s) of the search by choosing $K \in \{1, 2, 3\}$. The system can process queries across multiple layers simultaneously to provide a comprehensive set of results. The

search process in our system remains basic, focused, and represents events in natural language.

Users can enter prompts, adjust the abstraction level by switching to the corresponding tabs (Figure 2). Shot search ($K = 1$) performs a standard nearest neighbor search on each individual image embedding, ideal for specific queries.

In contrast, shot ($K = 2$) or action ($K = 3$) level search uses a two-stage retrieval process designed for efficiency and relevance:

1. **Segment Retrieval:** The user query is first compared to a compact set of high-level shot or action embeddings. This step quickly identifies the most relevant shots or events in the entire dataset, effectively reducing the search space exponentially.
2. **Detailed Selection:** For each top-ranked group (shot or action) retrieved in the first phase, the system performs a second local search. The system compares the user's query with Level 1 embeddings of only the images in that group. The single best-matching image from the group is then selected as a representative image.

Finally, the system displays these representative images as thumbnails in the user interface. A single thumbnail can represent an entire cohesive shot or a complex event, allowing the user to quickly browse high-level concepts. This hierarchical approach transforms the search experience from searching individual temporal units to optimally balancing retrieval efficiency and semantic accuracy

4 Conclusion

This work introduces a Hierarchical approach to interactive video retrieval by replacing the traditional flat index model with a semantically structured hierarchical architecture. By decomposing data representations into image, shot, and action levels, we demonstrate that the inherent trade-off between retrieval efficiency and depth of semantic understanding can be resolved, providing users with more contextually relevant and coherent results. In essence, we have shifted the problem from searching a large vector space to having users decide which vector space best fits concepts at different levels of abstraction.

However, the current architecture is just the beginning. We see event relational Modeling as a promising direction. There is a need to develop methods to explicitly model complex causal, temporal, and dependency relationships between sub-events, possibly using relational representation structures within events.

Acknowledgment. This publication has emanated from research conducted with the financial support of or supported in part by a grant from Science Foundation Ireland under Grant numbers 18/CRT/6223 and 13/RC/2106_P2 at the ADAPT SFI Research Centre at Dublin City University and the support of the Faculty of Engineering & Computing, DCU.

References

1. Cheng, Y.T., Wu, J., Ma, Z., He, J., Wei, X.Y., Ngo, C.W.: Interactive video search with multi-modal llm video captioning. In: Ide, I., Kompatsiaris, I., Xu, C., Yanai, K., Chu, W.T., Nitta, N., Riegler, M., Yamasaki, T. (eds.) MultiMedia Modeling. pp. 302–309. Springer Nature Singapore, Singapore (2025)
2. Dave, I.R., Heilbron, F.C., Shah, M., Jenni, S.: Sync from the sea: Retrieving alignable videos from large-scale datasets (2024), https://arxiv.org/abs/2409.01445
3. Gia, B.T., Khanh, T.B.C., Thanh, T.L.T., Doan, T.T., Le, K., Do, T., Mai, T.D., Ngo, T.D., Le, D.D., Satoh, S.: Nii-uit at vbs2025: Multimodal video retrieval with llm integration and dynamic temporal search. In: Ide, I., Kompatsiaris, I., Xu, C., Yanai, K., Chu, W.T., Nitta, N., Riegler, M., Yamasaki, T. (eds.) MultiMedia Modeling. pp. 318–325. Springer Nature Singapore, Singapore (2025)
4. Ho-Le, M.Q., Ho, D.K., Do-Huu, H.H., Le-Hinh, N.T., Vo-Hoang, H.V., Ninh, V.T., Gurrin, C., Tran, M.T.: Snapseek 2.0 at video browser showdown 2025. In: Ide, I., Kompatsiaris, I., Xu, C., Yanai, K., Chu, W.T., Nitta, N., Riegler, M., Yamasaki, T. (eds.) MultiMedia Modeling. pp. 339–346. Springer Nature Singapore, Singapore (2025)
5. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision (2021), https://arxiv.org/abs/2102.05918
6. Khan, O.S., Zhu, H., Sharma, U., Kanoulas, E., Rudinac, S., Jónsson, B.: Exquisitor at the video browser showdown 2024: Relevance feedback meets conversational search. In: Rudinac, S., Hanjalic, A., Liem, C., Worring, M., Jónsson, B., Liu, B., Yamakata, Y. (eds.) MultiMedia Modeling. pp. 347–355. Springer Nature Switzerland, Cham (2024)
7. Le-Hinh, N.T., Huynh, C.T., Ho-Le, M.Q., Ho, D.K., Tran, M.T., Huynh, V.T.: Lumina-1: Learning and understanding multimedia in immersive navigable archives for lifelog retrieval. In: Proceedings of the 8th Annual ACM Workshop on the Lifelog Search Challenge. p. 15–22. LSC '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3729459.3748698, https://doi.org/10.1145/3729459.3748698
8. Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., Ping, W.: Nv-embed: Improved techniques for training llms as generalist embedding models. arXiv preprint arXiv:2405.17428 (2024), https://arxiv.org/abs/2405.17428
9. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023), https://arxiv.org/abs/2301.12597
10. Nguyen, T., Anh, V.N.M., Pham, D.D., Vinh, T.Q., Quynh, N.D.T., Tien, L.A., Le, T.D., Nguyen, B.T.: Horus: Multimodal large language models framework for video retrieval at vbs 2025. In: Ide, I., Kompatsiaris, I., Xu, C., Yanai, K., Chu, W.T., Nitta, N., Riegler, M., Yamasaki, T. (eds.) MultiMedia Modeling. pp. 286–293. Springer Nature Singapore, Singapore (2025)
11. Nguyen-Ho, T.L., Huynh, V.T., Kongmeesub, O., Tran, M.T., Nie, D., Healy, G., Gurrin, C.: Veagle: Eye gaze-assisted guidance for video browser showdown. In: Ide, I., Kompatsiaris, I., Xu, C., Yanai, K., Chu, W.T., Nitta, N., Riegler, M., Yamasaki, T. (eds.) MultiMedia Modeling. pp. 347–354. Springer Nature Singapore, Singapore (2025)

12. OpenAI, Achiam, J., et al.: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774
13. Pantelidis, N., Pegia, M., Galanopoulos, D., Apostolidis, K., Stavrothanasopoulos, K., Moumtzidou, A., Gkountakos, K., Gialampoukidis, I., Vrochidis, S., Mezaris, V., Kompatsiaris, I., Jónsson, B.: Verge in vbs 2024. In: Rudinac, S., Hanjalic, A., Liem, C., Worring, M., Jónsson, B., Liu, B., Yamakata, Y. (eds.) MultiMedia Modeling. pp. 356–363. Springer Nature Switzerland, Cham (2024)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
15. Schoeffmann, K.: Video browser showdown 2012-2019: A review. In: 2019 International Conference on Content-Based Multimedia Indexing (CBMI). pp. 1–4 (2019). https://doi.org/10.1109/CBMI.2019.8877397
16. Schoeffmann, K., Ahlström, D., Bailer, W., Cobârzan, C., Hopfgartner, F., McGuinness, K., Gurrin, C., Frisson, C., Le, D.D., Del Fabro, M., Bai, H., Weiss, W.: The video browser showdown: A live evaluation of interactive video search tools. International Journal of Multimedia Information Retrieval (MMIR) **3**, 113–127 (06 2014). https://doi.org/10.1007/s13735-013-0050-8
17. Schoeffmann, K., Lokoč, J., Bailer, W.: 10 years of video browser showdown. In: Proceedings of the 2nd ACM International Conference on Multimedia in Asia. MMAsia '20, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3444685.3450215
18. Schöffmann, K., Bailer, W.: Video browser showdown. SIGMultimedia Rec. **4**(2), 1–2 (jul 2012). https://doi.org/10.1145/2350204.2350205
19. Spiess, F., Gasser, R., Heller, S., Rossetto, L., Sauter, L., Schuldt, H.: Competitive interactive video retrieval in virtual reality with vitrivr-vr. In: Lokoč, J., Skopal, T., Schoeffmann, K., Mezaris, V., Li, X., Vrochidis, S., Patras, I. (eds.) MultiMedia Modeling. pp. 441–447. Springer International Publishing, Cham (2021)
20. Tran, Q.L., Nguyen, B., Jones, G.J.F., Gurrin, C.: Videoease at vbs2025: An interactive video retrieval system. In: Ide, I., Kompatsiaris, I., Xu, C., Yanai, K., Chu, W.T., Nitta, N., Riegler, M., Yamasaki, T. (eds.) MultiMedia Modeling. pp. 363–370. Springer Nature Singapore, Singapore (2025)
21. Vadicamo, L., Arnold, R., Bailer, W., Carrara, F., Gurrin, C., Hezel, N., Li, X., Lokoc, J., Lubos, S., Ma, Z., et al.: Evaluating performance and trends in interactive video retrieval: Insights from the 12th vbs competition. IEEE Access **12**, 79342–79366 (2024)
22. Vuong, G.H., Ho, V.S., Nguyen-Dang, T.T., Thai, X.D., Ho-Le, M.Q., Le, T.K., Pham, M.K., Ninh, V.T., Gurrin, C., Tran, M.T.: Viewsinsight2.0: Enhancing video retrieval for vbs 2025 with an automatic query generator powered by large language models. In: Ide, I., Kompatsiaris, I., Xu, C., Yanai, K., Chu, W.T., Nitta, N., Riegler, M., Yamasaki, T. (eds.) MultiMedia Modeling. pp. 371–377. Springer Nature Singapore, Singapore (2025)
23. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., Zhang, P.: Florence: A new foundation model for computer vision (2021), https://arxiv.org/abs/2111.11432
24. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023), https://arxiv.org/abs/2303.15343