Title	**Efficient Search and Browsing of Large-Scale Video Collections with Vibro**
Keyword	Content-based Video Retrieval · Exploration · Visualization · Image Browsing · Visual and Textual co-embeddings
Purpose/Output	Cải thiện giao diện người dùng (UI) để dễ tiếp cận hơn việc tìm kiếm tạm thời (temporal search), nâng cấp thuật toán shot-detection, thay thế keyword-based search bằng đầu vào đa dạng text và giảm thời gian truy vấn bằng cách áp dụng phương pháp graph-based approximate nearest neighbor.
Contribution	❖ **<u>Tiêu chí của những người chiến thắng trước đó:</u>** • Khai thác thứ tự thời gian của các cảnh khác nhau trong video, do đó việc hỗ trợ thời gian truy vấn là rất quan trọng. • Hệ thống cho phép xem nhiều hình ảnh đã giúp có được cái nhìn tổng quan hơn về kết quả và cho phép trải nghiệm duyệt web hiệu quả hơn [12]. • Co-embedding dữ liệu trực quan và văn bản [14] đã dẫn đến kết quả đặc biệt cho các tìm kiếm truy vấn văn bản. • Kích thước của dữ liệu lớn thì phương pháp approximate nearest neighbor giúp tăng đáng kể tốc độ tìm kiếm với một sự đánh đổi nhỏ về độ chính xác. ❖ **<u>Với 1 số tiêu chí trên, đã nâng cấp hệ thống trước đây ở 1 số khía cạnh:</u>** • Clean up UI (Hình 1) bằng cách loại bỏ một số thành phần hiếm khi được sử dụng và tối ưu hóa các thành phần khác để temporal search. • Giới thiệu 1 thuật toán shot-detection được cản tiến. • Thay thế cơ chế lan truyền keyword được sử dụng trước đây cho tìm kiếm văn bản bằng model CLIP, cho phép input đa dạng text so sánh với keyword-based search. • Cập nhật image graph được tối ưu hóa để approximate nearest neighbor search và học hiệu quả với các node lân cận.
Data source	❖ Dataset V3C1 [4] (chứa 7475 tệp video và 1000 giờ video content) được kết hợp với tập dữ liệu V3C2 (9760 video và 1300 giờ)
Input	❖ Đầu vào là rich text, image, namely sketch.
Method	❖ **<u>Cuộc thi Video Browser Showdown (VBS) có 3 task chính:</u>** • Tìm kiếm mục đã biết bằng hình ảnh (v-KIS – visual Known-Item Search) • Tìm kiếm mục đã biết bằng văn bản (t-KIS – text Known-Item Search) • Tìm kiếm video Ad-Hoc (AVS – Ad-Hoc Video search) Đối với cả hai nhiệm vụ KIS, mục tiêu là tìm một phân đoạn chính xác phù hợp với mô tả đã cho và đối với AVS, phải submit càng nhiều phân đoạn phù hợp càng tốt.

❖ **Video Preprocessing – Tiền xử lý video**

Tất cả các frame của video được tổng hợp thành 3 cấp độ trừu tượng:

- **Step 1**: chỉ 1 frame mỗi giây được lưu trữ, tất cả các frame khác đều được bỏ qua. Sau đó, các frame này được merge thành các keyframe bằng cách phân tích hình thức trực quan bằng một tính năng cấp thấp được làm thủ công, dựa trên color, edge histograms và phân tích tần số với tổng 50 dimensions.
- **Step 2:** frame đầu tiên được chọn làm keyframe. Tất cả frame tiếp theo được so sánh với frame trước đó và keyframe cuối cùng. Nếu cả 2 trường hợp trên, mức độ tương tự giảm xuống dưới ngưỡng, frame hiện tại sẽ trở thành keyframe mới. Quy trình này được lặp lại cho đến khi kết thúc video. Ngưỡng được đặt chỉ để lọc ra các khóa gần giống nhau.
- **Step 3:** các keyframe được nhóm thành các shot. Sử dụng kết hợp low-level features và high-level CNN features thu được từ ResNet152 với DARAC-Pooling. Đối với V3C1 dataset, bao gồm 7475 video với tổng thời lượng 1000 giờ, việc xử lý trước này tạo ra hơn 3,5 triệu frames, 1,3 triệu keyframes và 700k shot (tiết kiệm khoảng 30% shot so với dataset được cung cấp ban đầu)

❖ **Navigation and Visualization**

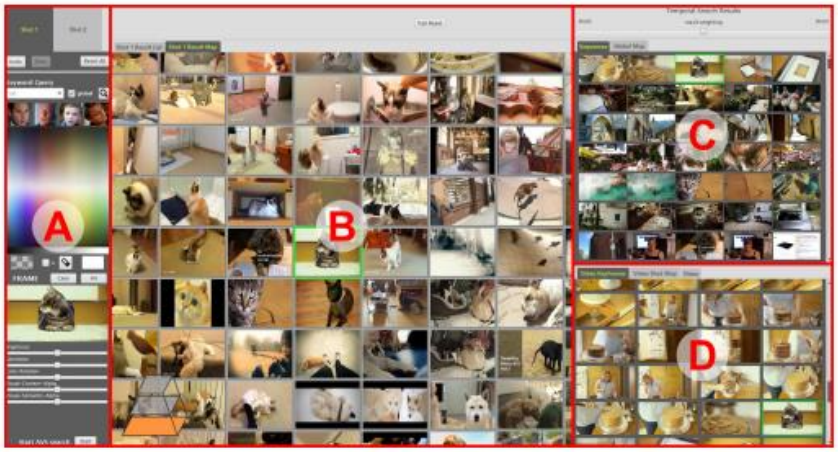

Fig. 1: Revised user interface for the current version of *Vibro*

- **Phần A:** được dành riêng cho việc xây dựng truy vấn.
- **Phần B:** 4000 kết quả hàng đầu của truy vấn hiện tại được hiển thị trong bản đồ được sắp xếp 2D bằng cách áp dụng tối ưu hóa SSM [3] (mất 1 phần nhỏ của giây). Các truy vấn tìm kiếm có thể được lập công thức cho hai ảnh khác nhau (các nút tab ở đầu phần A) và kết quả hiển thị thay đổi khi các ảnh này được chuyển đổi. Hơn nữa, hai truy vấn này được kết hợp trong một temporal search và kết quả của chúng được trình bày trong phần C dưới dạng danh sách five Efficient Search và

Browsing of Large-Scale Video Collections với các shot-frames liên tục của Vibro.

- Video chứa frame đã chọn (viền xanh lục) được hiển thị trong phần D bằng trình phát video hoặc danh sách tất cả các keyframe.

➔ Hệ thống hoạt động khá tốt trong task KIS nhưng lại quá chậm với task AVS, vì thế đã thiết kế 1 UI riêng cho phần này. Sau khi chọn một hình ảnh từ main UI, image-based search được kích hoạt bởi nút AVS "Start" ở cuối phần A. 20 hình ảnh tốt nhất được hiển thị trên một UI riêng biệt. Bây giờ người dùng đánh dấu tất cả positive frames và gửi chúng đến máy chủ đánh giá. Hơn nữa, các frame đã gửi được sử dụng trong một vòng phản hồi về mức độ liên quan. Cùng với hình ảnh ban đầu và các positive frames khác, chúng tạo thành một truy vấn nhiều hình ảnh để tạo ra 20 kết quả tiếp theo.

❖ **Search Modalities – Phương thức tìm kiếm:**

Vibro cho phép các truy vấn theo ba phương thức, đó là namely sketch, text và hình ảnh theo ví dụ. Ngoài ra, kết quả tìm kiếm của ba phương pháp này được kết hợp với các trọng số có thể điều chỉnh thành một danh sách kết quả duy nhất, sau đó được hiển thị trên UI.

- *Sketches and Images*: Trong tìm kiếm dựa trên bản phác thảo, các truy vấn có thể được xây dựng bằng cách vẽ trên canvas trống hoặc bằng cách chọn một hình ảnh mẫu từ bất kỳ phần nào của UI và sửa đổi giao diện của nó bằng các công cụ vẽ được cung cấp. Để kích hoạt chức năng này, trích xuất hai vectơ đặc trưng khác nhau từ tất cả các keyframe. Low-level feature vector đầu tiên (đã được sử dụng để tiền xử lý video) và 768-dim high-level CNN feature vector thứ hai thu được từ ResNetx16 được sử dụng trong phần trực quan của mô hình CLIP. Feature đầu tiên được sử dụng để có được hình thức trực quan (bố cục màu sắc và hình dạng) và được tính toán lại nếu hình ảnh bị sửa đổi, trong khi feature thứ hai mô tả nội dung ngữ nghĩa và chỉ được tính toán một lần.
- *Rich Text:* Đầu vào tìm kiếm thứ ba có thể là văn bản. Với cách tiếp cận sử dụng cơ chế lan truyền keyword, thông tin về sự tương tác giữa các keyword riêng lẻ bị thiếu. Để khắc phục hạn chế này, nhóm tác giả đã tích hợp CLIP – một mô hình co-embedding mã hóa đầu vào văn bản đa dạng và image vào một không gian vectơ dùng chung.
- *Sự kết hợp của các phương thức:* Vì low-level feature chỉ có 50-dims nên có thể thực hiện tìm kiếm tuyến tính trong thời gian dưới 100 ms trên toàn bộ dataset, thu

được similarity score cho từng keyframe và ghi nhớ giá trị của chúng để kết hợp potential score. Truy vấn text và image được tăng tốc bởi thuật toán Graph-based approximate nearest neighbor search, mỗi truy vấn trả về 4000 kết quả được sắp xếp theo similarity score giảm dần. Vì text và image features có chung không gian nhúng nên chỉ cần 1 graph. Để kết hợp 2 kết quả tìm kiếm low và high-level visual features, sử dụng trọng số trung bình của các similarity score trong đó weight ban đầu $\alpha \in [0, 1]$ được đặt = 0.5 và có thể thay đổi. Nếu truy vấn văn bản được sử dụng với truy vấn phác thảo hoặc truy vấn hình ảnh, thực hiện tìm kiếm dựa trên graph-based với text embedding và kết hợp similarity score với weight có thể điều chỉnh $\beta \in [0, 1]$.

- *Temporal Queries:* Đối với hai dấu thời gian khác nhau, có thể thực hiện bất kỳ truy vấn đa phương thức nào được mô tả và thu được similarity score của các keyframe cho từng truy vấn đó. Để cho phép tìm kiếm tạm thời, xếp hạng các chuỗi keyframe liên tiếp từ một video theo xác suất Efficient Search và Browsing of Large-Scale Video Collections mà chuỗi chứa nội dung từ tab truy vấn đầu tiên, tiếp theo là nội dung từ tab thứ hai. Similarity score từ dấu thời gian đầu tiên của mỗi khung được kết hợp với điểm cao nhất của dấu thời gian thứ hai cho tất cả các khung tiếp theo trong một khoảng thời gian nhất định. Kết quả cuối cùng sau đó được sắp xếp theo các tổng similarity scores này.

❖ **Graph-Based Nearest Neighbor Search**

- Thời gian cần thiết để tìm kiếm toàn bộ bộ sưu tập hình ảnh phụ thuộc vào số lượng hình ảnh và độ phức tạp của phép tính độ tương đồng giữa truy vấn và hình ảnh cơ sở dữ liệu. Cái sau bị ảnh hưởng bởi kích thước của vectơ đặc trưng và công thức tính độ tương tự. Vì các tính năng không thể được nén vô hạn và số lượng hình ảnh trong hầu hết các tập dữ liệu không ngừng tăng lên, nên đã hướng tới *approximate nearest neighbor searches*. Các phương pháp như vậy không tính đến tất cả dữ liệu; tùy thuộc vào thuật toán, một số tính năng hình ảnh bị lược bỏ hoặc heavily quantized. Kết quả không chính xác được đổi lấy để tăng tốc độ tính toán lên một hoặc hai độ lớn.
- Phương pháp trước đây là giảm các CNN feature mà không PCA whitening xuống còn 64-dims và định lượng theo bytes. Các feature này hoạt động kém hơn gần 10% so với high-dimensional features với PCA whitening trong hệ thống content-

	based image retrieval, tuy nhiên high-dimensional features lại không thể nén và không thể được sử dụng trong một hệ thống tương tác vì mất quá nhiều thời gian để tìm kiếm hàng triệu hình ảnh. Vì thế, nhóm tác giả đã thiết kế lại cấu trúc dữ liệu graph cho kết quả tìm kiếm hàng xóm gần nhất gần đúng.
Result	❖ Sau khi thiết kế lại, mô hình mới cho các kết quả giống nhau gần 99% và nhanh hơn 20 lần. Graph hiệu quả hơn so với các phương pháp SOTA khác, yêu cầu ít bộ nhớ hơn và cung cấp hiệu quả hơn các vùng lân cận tương tự để khám phá tương tác.
Limitation	
Future research	
New idea	
Reference	[4] Berns, F., Rossetto, L., Schoeffmann, K., Beecks, C., Awad, G.: V3c1 dataset: An evaluation of content characteristics. In: ICMR '19 Proceedings of the 2019 on International Conference on Multimedia Retrieval (2019) [9] Heller, S., Gasser, R., Illi, C., Pasquinelli, M., Sauter, L., Spiess, F., Schuldt, H.: Towards explainable interactive multi-modal video retrieval with vitrivr. In: Lokoˇc, J., Skopal, T., Schoeffmann, K., Mezaris, V., Li, X., Vrochidis, S., Patras, [12] Kratochv´ıl, M., Vesel`y, P., Mejzl´ık, F., Lokoˇc, J.: Som-hunter: Video browsing with relevance-to-som feedback loop. In: International Conference on Multimedia Modeling. pp. 790–795. Springer (2020) I. (eds.) MultiMedia Modeling. pp. 435–440. Springer International Publishing, Cham (2021) [14] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. CoRR abs/2103.00020 (2021)