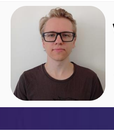
Experimental Evaluation of Static Image Sub-Region Based Search Models Using CLIP


Bastian Jäckl
bastian.jaeckl@uni-konstanz.de

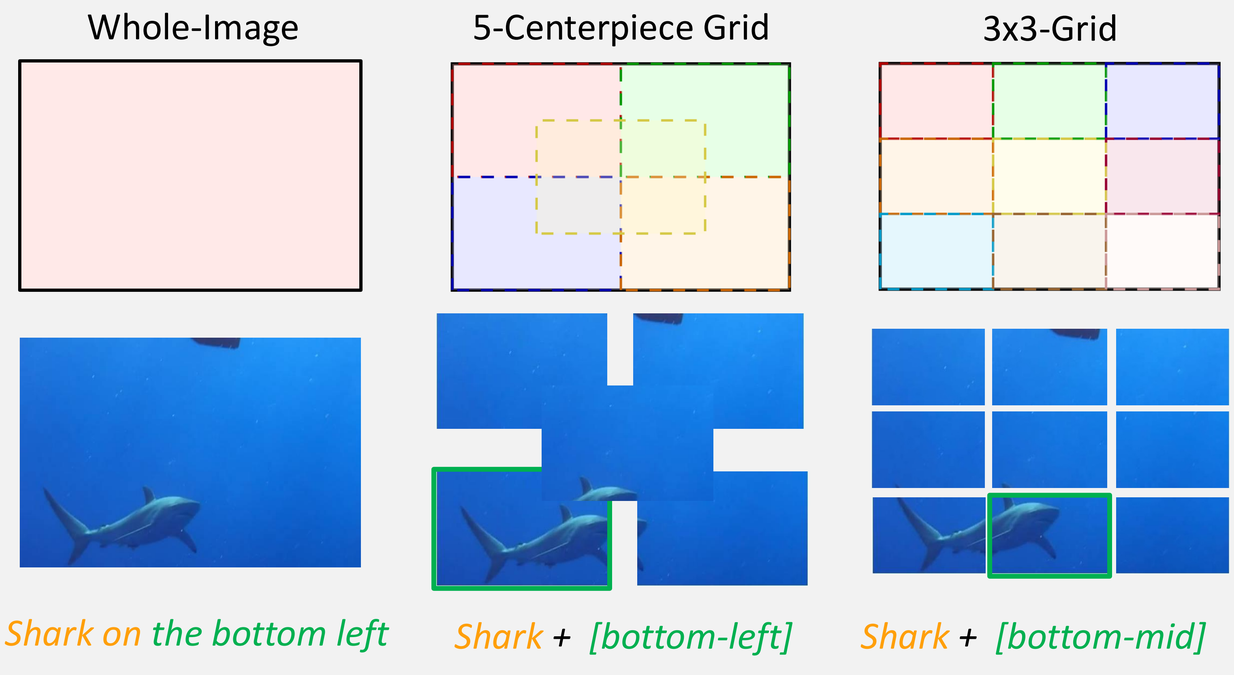

Vojtěch Kloda
vojtech.kloda825@student.cuni.cz


Daniel A. Keim
keim@uni-konstanz.de


Jakub Lokoč
jakub.lokoc@matfyz.cuni.cz

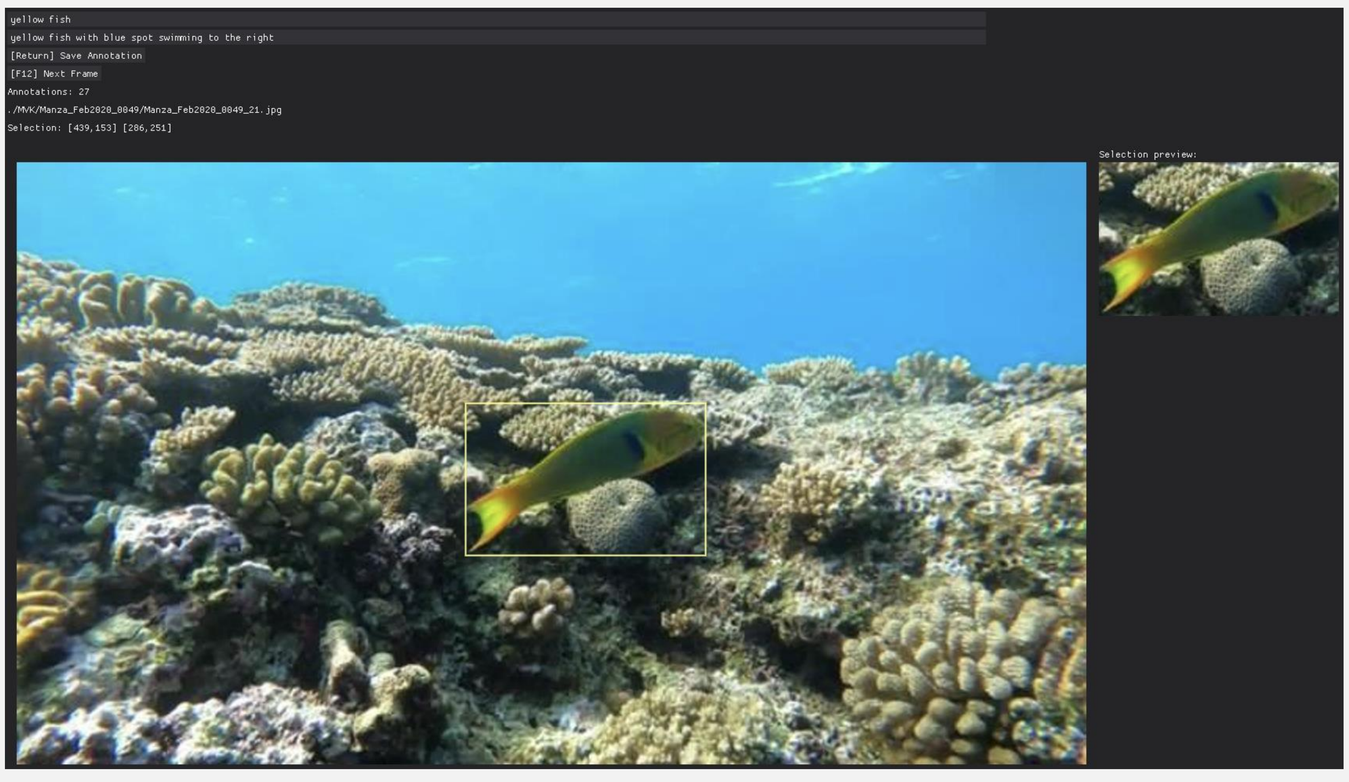
1) Motivation

- Multimodal retrieval systems enable users to **query** for images/videos **by text** [1].
- Static sub-region search extends **semantic text queries** with **position information** specifying which sub-region the query should match.
- Used by interactive retrieval systems in competitions such as the Video Browser Showdown [2].
- **Experimental evaluation** of the effectiveness of static grids **is missing.**



2) Data Collection

- Collected **741 annotations** of images from **12 annotators.**
- Each annotation contains a **long description, a short description,** and the **bounding box** where the descriptions should match.
- **Two types** of annotations: **Skippable** (where annotators could skip images that they **did not want to annotate**) and **Unskippable** annotations.
- We chose the challenging **MVK dataset** [3] containing more than 28 hours (84309 keyframes) of underwater content.
- Annotators were able to draw bounding boxes directly into the images. Consequently, the experiments show **theoretical upper bounds** utilizing perfect bounding boxes.



3) Evaluation and Results

- Investigate **retrieval effectiveness** of **whole-image baseline, two static grid configurations, and theoretical oracle** that estimates the potential of region-based retrieval with perfect bounding boxes.
- **Grids beat whole-image:** Both the **5-centerpiece grid** and the **3x3 grid** consistently improve over **whole-image** CLIP.
- Gains come from **better localization** when a cell overlaps the target and **noise filtering** that limits irrelevant areas from affecting the embedding.
- **Textual position alone fails.** Adding phrases like “top left” to the query even **reduces** effectiveness. Positional intent must be enforced **spatially**, not just linguistically.
- **Robustness & ceiling.** Moderate **overlap (10%)** of grid partitionings exhibits the highest effectivenss and mitigates shift/size noise.
- Even the **theoretical oracle** crop caps around **R@100 ≈ 41/30%** (skipbbable/unskippable queries).

4) Conclusion

- **Take-Home Message:** Lightweight **static sub-region search** consistently beats **whole-image search** in homogeneous imagery.
- **Limitations:** Results use **perfect annotation boxes** for queries and were evaluated on **a single dataset.**
- **Future Work:** Improve region embeddings by leveraging **spatially coherent regions** (beyond fixed grids) to tighten localization and prefiltering, and assess the gains under **imperfect boxes** and across **different datasets**.

[1] Radford et al.: Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning, 2021.
[2] Stroh et al.: Prak Tool v3: Enhancing Video Item Search Using Localized Text and Texture Queries International Conference on Multimedia Modeling, 2025.
[3] Truong et al.: Marine Video Kit: A New Marine Video Dataset for Content-Based Analysis and Retrieval. International Conference on Multimedia Modeling, 2023.

Bastian Jäckl[1], Vojtěch Kloda[2], Daniel A. Keim[1], Jakub Lokoč[2]
1. Data Analysis and Visualization Group, University of Konstanz, Germany
2. SIRET Research Group, Charles University, Czechia

October 01-03, 2025, Reykjavík, Iceland