Marine Video Kit: A New Marine Video Dataset for Content-based Analysis and Retrieval

Quang-Trung, Truong[1], Tuan-Anh Vu[1], Tan-Sang Ha[1], Jakub Lokoč[2], Yue Him Wong Tim[3], Ajay Joneja[1], and Sai-Kit Yeung[1]

[1] Hong Kong University of Science and Technology
[2] FMP, Charles University, Prague, Czech Republic
[3] Shenzhen University

Abstract. Effective analysis of unusual domain specific video collections represents an important practical problem, where state-of-the-art general purpose models still face limitations. Hence, it is desirable to design benchmark datasets that challenge novel powerful models for specific domains with additional constraints. It is important to remember that domain specific data may be noisier (e.g., endoscopic or underwater videos) and often require more experienced users for effective search. In this paper, we focus on single-shot videos taken from moving cameras in underwater environments, which constitute a nontrivial challenge for research purposes. The first shard of a new Marine Video Kit dataset is presented to serve for video retrieval and other computer vision challenges. Our dataset is used in a special session during Video Browser Showdown 2023. In addition to basic meta-data statistics, we present several insights based on low-level features as well as semantic annotations of selected keyframes. The analysis also contains experiments showing limitations of respected general purpose models for retrieval. Our dataset and code are publicly available at https://hkust-vgd.github.io/marinevideokit.

Keywords: Known-item Search · Underwater video · Dataset Statistics.

1 Introduction

In order to facilitate the development of multimedia retrieval and analysis models, the research community establishes and uses various benchmark multimedia datasets [9,22,24,14,23,2,9,22,24]. The datasets usually provide so-called ground-truth annotations and allow repeatable experimental comparison with state-of-the-art methods.

The source of large benchmark collections is often a video sharing platform (e.g., Youtube, or Vimeo) with specific licensing of the content. For example, the respected Vimeo Creative Commons collection dataset (V3C) [19] contains several thousand hours of videos downloaded from the Vimeo platform. Although only videos with the creative commons license were used for the dataset, it is available only with an agreement form indicating possible changes in the future.
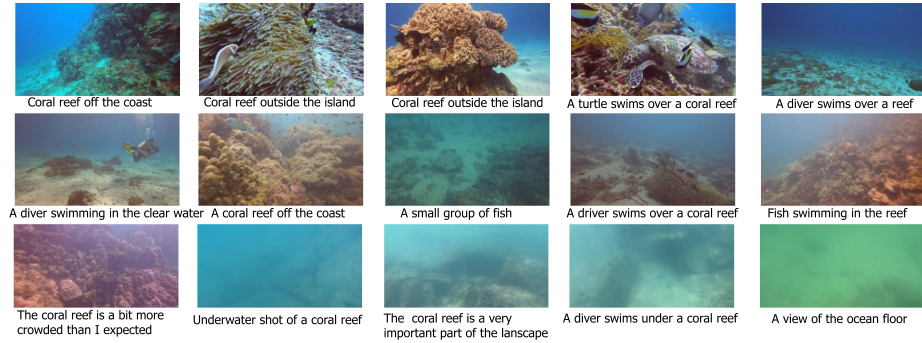
Coral reef off the coast	Coral reef outside the island	Coral reef outside the island	A turtle swims over a coral reef	A diver swims over a reef
A diver swimming in the clear water	A coral reef off the coast	A small group of fish	A driver swims over a coral reef	Fish swimming in the reef
The coral reef is a bit more crowded than I expected	Underwater shot of a coral reef	The coral reef is a very important part of the lanscape	A diver swims under a coral reef	A view of the ocean floor

Fig. 1. Several examples of dataset video frames and their ClipCap descriptions.

Hence, it is beneficial to design also new video datasets with a limited number of copyright owners, limiting potential future changes to the dataset (experiment repeatability). Designing datasets with highly challenging content is also necessary, even for a limited number of data items. This aspect is especially important for interactive search evaluation campaigns [7,8] addressing a broad community of researchers from different multimedia retrieval areas. Indeed, for many research teams (especially smaller ones), it might be more feasible to participate in a difficult challenge over 10-100 hours of videos, rather than a challenge over 10.000 hours and more.

In general everyday videos, there appear many common classes of objects, and thus even a larger collection can be effectively filtered with text queries. On the other hand, a domain specific collection (i.e., one cluster with a lower variance of common keywords) might already be challenging for lower sizes of collections. Therefore, we selected underwater marine videos with the seafloor, coral reefs, and various biodiversity where potentially effective keywords are unknown to ordinary users. Furthermore, unlike common "everyday" videos, underwater videos pose additional obstacles for multimedia analysis and retrieval models. These challenges include low visibility, blurry shots, varying sizes and poses of objects, a crowded background, light attenuation, and scattering, among others [10]. Therefore, not only general purpose models but also domain specific classifiers require novel ideas and breakthroughs to reach human-level accuracy.

This paper presents a first fragment of a new "Marine Video Kit" dataset, currently intended mainly for content-based retrieval challenges. It is composed of more than 1300 underwater videos from 36 locations worldwide and at different times across the year. Whereas the long-term ambition for this dataset might be even a video-sharing platform with controllable extensions, for now, we present a manually organized set of directories comprising videos, selected frames, and various forms of meta-data. So far, the meta-data comprises available video attributes (location, time) and pre-computed captions as well as embeddings for

the set of selected video frames. In the future, we plan to provide also more annotations for object detection, semantic segmentation, object tracking, etc.

The remainder of this paper is structured as follows: section 2 gives an overview of related works, section 3 introduces the details of the Marine Video Kit dataset, and finally, section 4 concludes the paper.

2   Related Work

Numerous datasets were created for the object detection and segmentation task in order to better comprehend marine life and ecosystems. In this section, we briefly review some recent works for Marine-related datasets.

The Brackish [16] is an open-access underwater dataset containing annotated image sequences of starfish, crabs, and fish captured in brackish water with varying degrees of visibility. The videos were divided into categories according to the primary activity depicted in each one. A bounding box annotation tool was then used to manually annotate each category's 14,518 frames, producing 25,613 annotations in total.

MOUSS dataset [4] is gathered by a horizontally-mounted, grayscale camera that is placed between 1 and 2 meters above the sea floor and is illuminated solely by natural light. In most cases, the camera remains stationary for 15 minutes at a time in each position. There are two sequences in the MOUSS datasets: MOUSS seq0 and MOUSS seq1. The MOUSS seq0 includes 194 images, all of which belong to the Carcharhini-formes category, and each image has a resolution of 968 by 728 pixels. There is only one category in the MOUSS seq1, which is called Perciformes. Each image has a resolution of 720 by 480 pixels. A human expert was responsible for assigning each of the species labels.

WildFish [25] is a large-scale benchmark for fish recognition in the wild. It consists of 1,000 fish categories and 54,459 unconstrained images. This benchmark was developed for the field of the classification task. In the field of image enhancement, the database known as Underwater ImageNet [5] is made up of subsets of ImageNet [3] that contain photographs taken underwater. As a result, the distorted and undistorted sets of underwater images were allowed to have 6,143 and 1,817 pictures, respectively. Fish4Knowledge [6] gave an analysis of fish video data.

OceanDark [21] is a novel low-lighting underwater image dataset that was created to quantitatively and qualitatively evaluate the proposed framework. This dataset was developed in the field of image enhancement and consisted of images that were captured using artificial lighting sources.

The Holistic Marine Video Dataset [11], also known as HMV, is a long video that simulates marine videos in real time and annotates the frames of HMV with scenes, organisms, and actions. The goal of this dataset is to provide a large-scale video benchmark with multiple semantic aspect annotations. On the other hand, they also provide baseline experiments for reference on HMV for three tasks: the detection of marine organisms, the recognition of marine scenes, and the recognition of marine organism actions.

3   Marine Video Kit Dataset

In this section, we will show details about our Marine Video Kit Dataset, which provides a dense, balanced set of videos focusing on marine environment, hence enriching the pool of existing marine dataset collections.

Many types of cameras were used to build the presented first shard of the Marine Video Kit dataset, such as Canon PowerShot G1 X, Sony NEX-7, OLYM-PUS PEN E-PL, Panasonic Lumix DMC-TS3, GoPro cameras, and consumer cellphones cameras. The dataset consists of a larger number of single-shot videos without post-processing. Unlike common video collections taken by crawling search engines, the presented dataset focuses solely on marine organisms captured during diving periods. The typical duration of each video is about 30 seconds.

To illustrate the specifics of the underwater environment, we present 3D color histograms showing differences for different parts of the dataset. For a larger set of selected frames (extracted with 1fps), we also analyze semantic descriptions automatically extracted by the ClipCap model [15]. For a randomly selected subset of 100 frames, the automatically generated descriptions are compared with manually created annotations in a known-item search experiment.

Fig. 2. The world map illustrates countries/regions where we capture data around the world (map source: Google Maps).

3.1   Acquiring the Dataset

Different ocean areas have their own marine biodiversities, such as various animals, plants, and microorganisms. Therefore, to create a diverse dataset for the

research community, we have visited and captured data from 11 different regions and countries during daytime and nighttime (unless bad weather or other circumstances happen). We present a world map illustrating 11 countries/regions where the videos were captured Fig. 2.

We categorize the captured videos in terms of their location and time, then utilize OpenCV library [1] to process all data into one unifying format (JPG for images, MP4 for videos). Due to the variety of capturing devices, our raw videos have different resolutions, from the HD (720p) resolution to Ultra HD (4K), with a frame rate of 30 fps. Therefore, we also utilize FFmpeg library[20] to convert all data to low and high resolution for different research purposes, such as video retrieval, super-resolution, object detection, segmentation, etc.

3.2 Dataset Structure
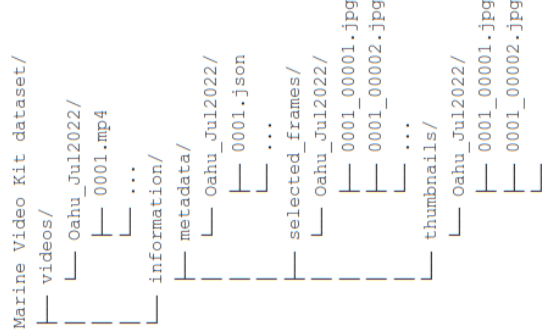
Fig. 3. Directory structure of the Marine Video Kit dataset.

In this section, we show our data's directory structure, which organizes different aspects of the data. As shown in Fig. 3, there are two sub-directories for video and its supplementary information.

For each video directory, we format their name as *location_time* pattern to explicitly represent the time and location that they were captured. For example, "Oahu_Jul2022" was captured at Oahu - the third-largest of the Hawaiian Islands, in July 2022.

For each information directory, the *selected_frames* directory stores all frames that are evenly selected one frame per second and are kept in the original resolution, while the *thumbnails* directory stores the same frame but in down-scaled resolution. Finally, the *metadata* directory contains the associated meta information of each video in JSON format for easier sharing and parsing information. This metadata file contains semantic and statistical information, such as video name, duration, height, width, camera device, directory, license, and reference information, such as ClipCap captions.
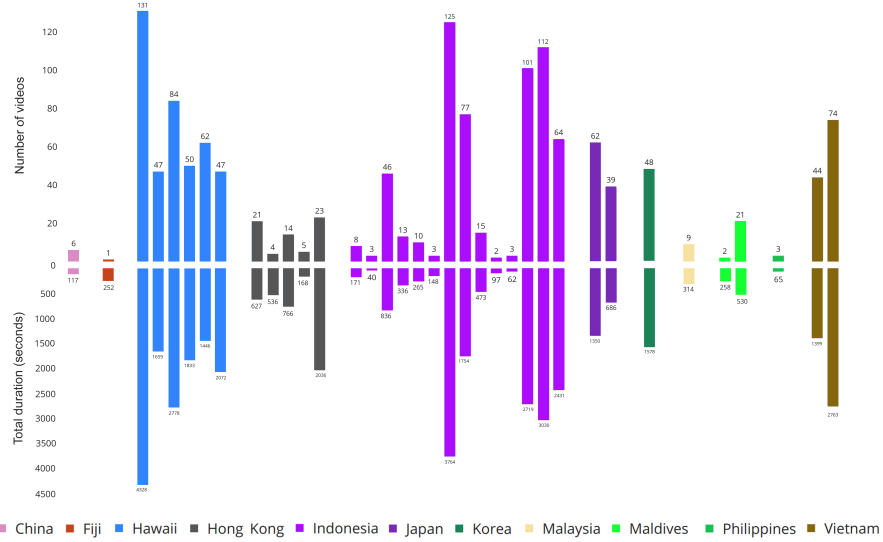
3.3   Dataset statistics

Fig. 4. The figure shows the number of videos and overall time duration for each region.

We capture data from 11 different regions and countries during the time from 2011 to 2022. There is a total of 1379 videos with a length from 2 seconds to 4.95 minutes, with the mean and median duration of each video is 29.9 seconds, and 25.4 seconds, respectively. The total duration is slightly above 12 hours, however, the diving time is significantly larger, up to a thousand hours. Fig. 4 shows the number of videos, and the total length of videos varies by category.

To represent marine videos from directories in color space, 3D color histograms are used to aggregate pixels to a $256 \times 256 \times 256$ array for 3D representation. The 3D visualization of each directory is presented in Fig. 5, 6 with 3D points colored by RGB histogram values based on corresponding places in 3D spaces, and the sizes are proportional to the color densities. We used color quantization and normalization for a more informative illustration of 3D color histogram densities.

A big challenge when using Marine Video Kit dataset for retrieval tasks is that the marine environment has changed considerably, and the captured data is heavily affected by lighting variations or caustics, thus causing low visibility. Fig. 5 illustrates regions under low lighting, and Fig. 6 mentions regions under good lighting.
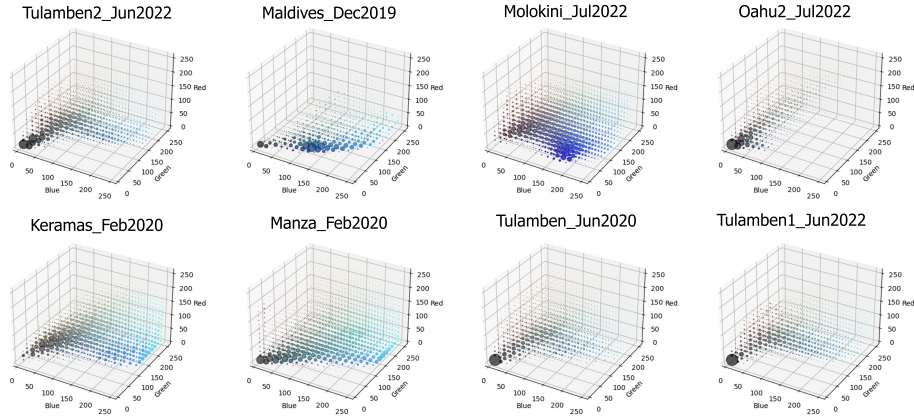
Fig. 5. Illustration of video frame colors present in regions under low illumination using 3D color histograms.

3.4 ClipCap Captions

To provide semantic information for the set of uniformly selected frames, ClipCap [15] architecture was employed. ClipCap elegantly combines rich features of the CLIP model [18] with the powerful GPT-2 language model [12]. Specifically, a prefix is computed for the image feature vector, while the language model continues to generate the caption text based on the prefix. Additionally, ClipCap is efficient in diverse datasets, which motivates us to caption selected frames in the marine dataset. Captions given by ClipCap are utilized as an automatically generated caption attribute in video metadata files.

We adopt ClipCap for selected frames in Fig. 7 that are mentioned in metadata files. The process to generate captions consists of two steps to ensure content relevance. The captions of selected frames are automatically outputted by the ClipCap model; after that, we regularly inspect the captions to remove unrelated descriptions. Fig. 1 shows selected frame captions of ClipCap that are utilized to describe semantic information of the marine dataset, and Fig. 8 presents frequencies for individual words in frame captions.

3.5 Known-item Search Experiment

In this section, we show that the new dataset represents a challenge for an information retrieval task. Specifically, a known-item search experiment is performed using a subset of all selected video frames O_i (extracted with 1fps, about 40K frames), and their CLIP representations [17]. The experiment assumes a set of pairs $[q_i, t_i]$, where t_i is a CLIP embedding vector representing a target (searched) video frame O_i, and q_i is a CLIP embedding vector representing a query description of the target image O_i. Using one pair, it is possible to rank
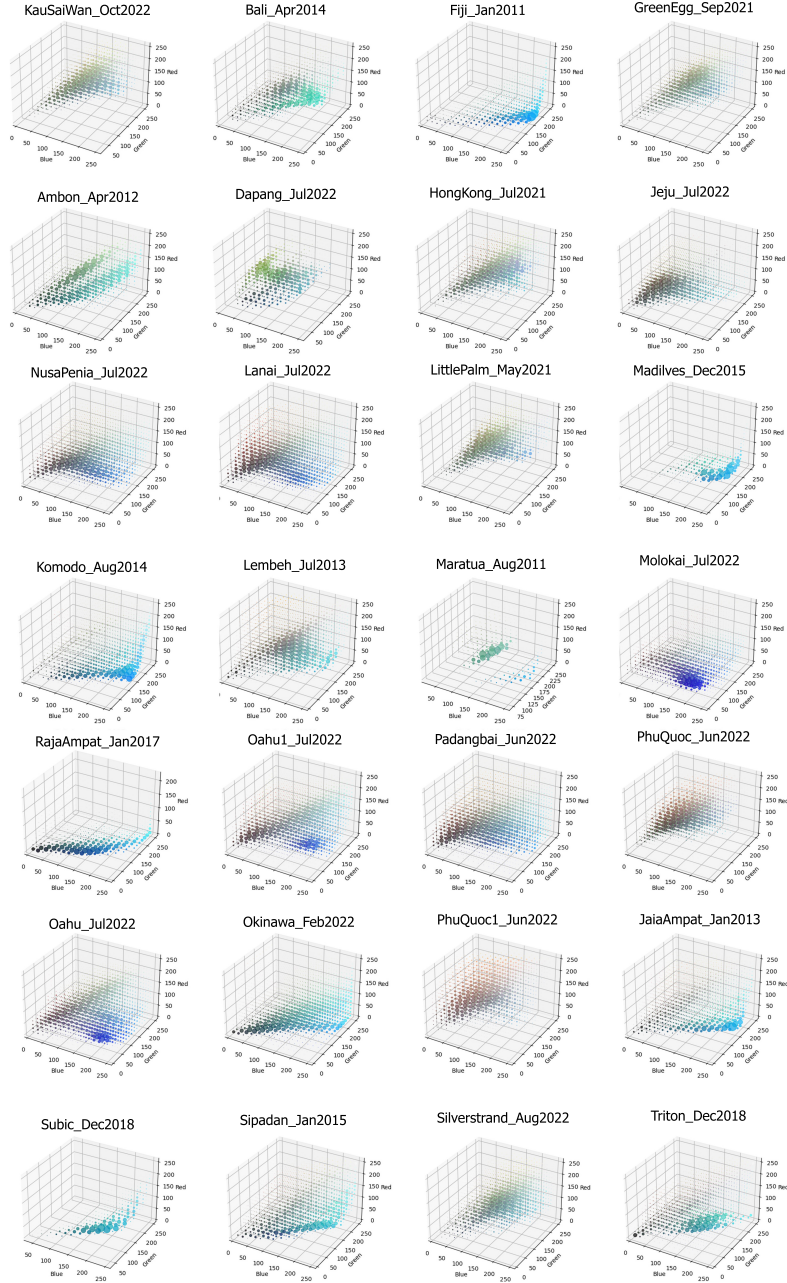
Fig. 6. Illustration of video frame colors present in regions under good illumination using 3D color histograms.
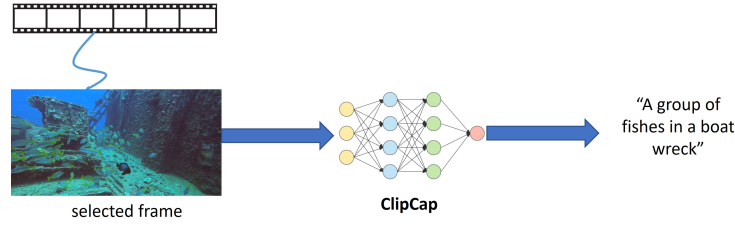
Fig. 7. The captioning architecture. A selected frame is extracted from marine video, then fed into the ClipCap model to output the caption of the selected frame.

all selected video frames based on their cosine distance to the query vector q_i. From this ranking, the rank of the target image O_i can be identified and stored. Repeating this experiment for all available pairs $[q_i, t_i]$, a distribution of ranks of searched items can be analyzed.

Figure 9 shows the result of our preliminary KIS experiment, where 100 pairs $[q_i^{novice}, t_i]$ and 100 pairs $[q_i^{expert}, t_i]$ were created manually. Novice and VBS Expert annotators described randomly selected target images, where the VBS Expert is not an expert in the marine domain but has experience with query formulation from the Video Browser Showdown. ClipCap annotations were computed for the same target images as well. Using CLIP embeddings, ranks of target images were computed for the novice, VBS Expert, and ClipCap queries. Figure 9 shows that the overall distributions of ranks are similar for ClipCap and novice users, while the VBS expert was able to reach a better average and median rank. In the future, we plan more thorough experiments. We also note that employed random selection of target images often does not lead to unique items, and users did not see the search results for the provided 100 text descriptions/queries. Both notes affect the overall distribution of ranks.
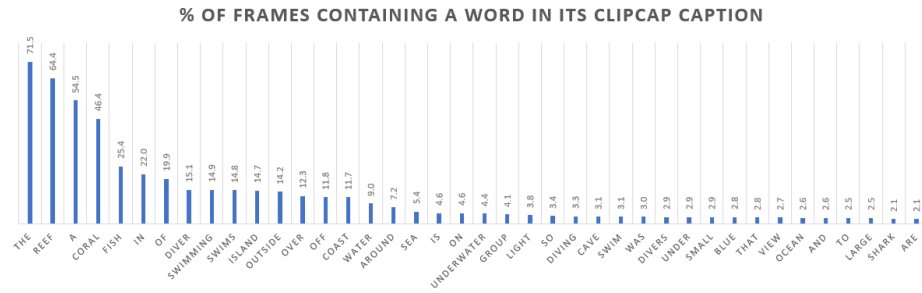

Fig. 8. Occurrence of words in frame captions. Computed for a subset of the dataset.
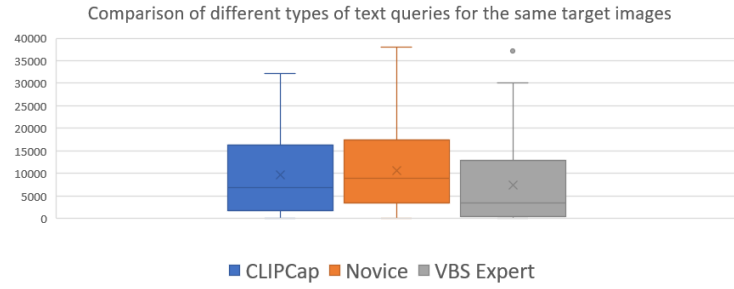
Fig. 9. Ranks for ClipCap, Novice, and VBS Expert text queries for 100 target images.

Since ClipCap descriptions are available for all database frames, we also present another KIS experiment for ClipCap based queries with similar average target rank as novice users in Figure 9. Figure 10 was evaluated for 4000 pairs $[q_i, t_i]$, where the target image descriptions were obtained using the ClipCap approach. We may observe that only about 30% of all queries are allowed to find the target in the top-ranked 2000 dataset items. Even for the small dataset. Although the CLIP retrieval model helps, it is indeed a way more difficult challenge than searching common videos with a large number of different concepts (e.g., compared to the results of a similar study here [13]). Furthermore, searching top 2000 items is not trivial. Looking at the top 500 close-ups of the histogram, only 6.6% of target items can be found in the top 100. Hence, we conclude that limited vocabulary (i.e., ClipCap like queries) and similar content make the known-item search challenge difficult for the proposed dataset even with the respected CLIP model.
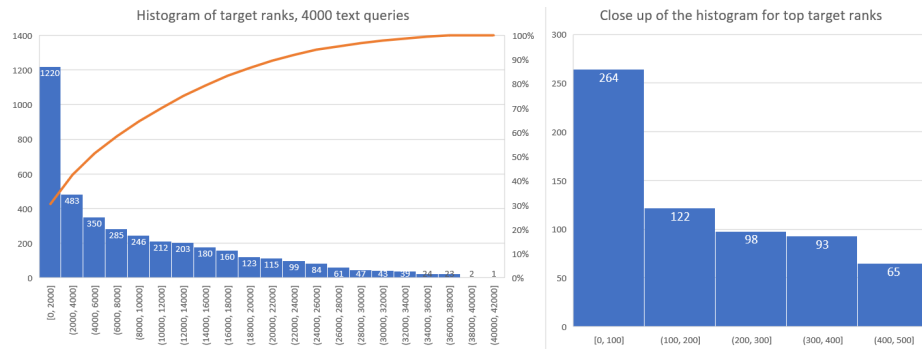

Fig. 10. 4000 KIS queries using ClipCap captions, histogram bins aggregate ranks of target images for the corresponding queries.

4 Conclusion

We provide the Marine Video Kit dataset, single-shot videos challenging for content-based analysis and retrieval. To provide a first insight of the new marine dataset, basic statistics based on meta-data, low-level color descriptors, and Clip-Cap semantic annotations are presented. We present a baseline retrieval study for the Marine Video Kit dataset to emphasize the domain's specificity. Our experiments show that the similarity of content in the dataset causes difficulties for a respected cross-modal based know-item search approach. We hope that our dataset with the baseline for content-based retrieval will accelerate considerable progress in the marine video retrieval area.

We plan to extend the dataset in the future with new annotations and videos from new environments. Besides, additional computer vision tasks over the dataset, such as semantic segmentation or object detection, could be prepared for the research community. Furthermore, motion analysis, fish counting, or detection tasks are also meaningful information for retrieval applications.

Disclaimer. Any opinions, findings, conclusions, or recommendations expressed in this material do not necessarily reflect the views of HKLTL, CAPCO, HK Electric, and the Marine Conservation Enhancement Fund.

References

1. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000) 5
2. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018) 1
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009) 3
4. Derya, A., Anthony, H., Suchendra, B.: Mouss dataset (2018) 3
5. Fabbri, C., Islam, M.J., Sattar, J.: Enhancing underwater imagery using generative adversarial networks. In: 2018 IEEE International Conference on Robotics and Automation (ICRA) (2018) 3
6. Fisher, R.B., Chen-Burger, Y.H., Giordano, D., Hardman, L., Lin, F.P., et al.: Fish4Knowledge: collecting and analyzing massive coral reef fish video data, vol. 104. Springer (2016) 3
7. Gurrin, C., Zhou, L., Healy, G., Jónsson, B.Þ., Dang-Nguyen, D., Lokoc, J., Tran, M., Hürst, W., Rossetto, L., Schöffmann, K.: Introduction to the fifth annual lifelog search challenge. In: International Conference on Multimedia Retrieval (2022) 2

8. Heller, S., Gsteiger, V., Bailer, W., Gurrin, C., Jónsson, B.Þ., Lokoč, J., Leibetseder, A., Mejzlík, F., Peška, L., Rossetto, L., Schall, K., Schoeffmann, K., Schuldt, H., Spiess, F., Tran, L., Vadicamo, L., Veselý, P., Vrochidis, S., Wu, J.: Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th video browser showdown. Int. J. Multim. Inf. Retr. (2022) 2

9. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: International Conference on Computer Vision (ICCV) (2017) 1

10. Levy, D., Levy, D., Belfer, Y., Osherov, E., Bigal, E., Scheinin, A.P., Nativ, H., Tchernov, D., Treibitz, T.: Automated analysis of marine video with limited data. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2018) 2

11. Li, Q., Li, J., Shi, Z., Gu, Z., Zheng, H., Zheng, B., Li, J.: A holistic marine video dataset. In: OCEANS 2021: San Diego – Porto (2021) 3

12. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021) 7

13. Lokoč, J., Souček, T.: How many neighbours for known-item search? In: Similarity Search and Applications - 14th International Conference, SISAP 2021 Proceedings (2021) 10

14. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: Proceeding of International Conference on Multimedia Retrieval (ICMR). ACM (2018) 1

15. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021) 4, 7

16. Pedersen, M., Haurum, J.B., Gade, R., Moeslund, T.B., Madsen, N.: Detection of marine animals in a new underwater dataset with varying visibility. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019) 3

17. 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 (2021) 7

18. 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) 7

19. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3c–a research video collection. In: International Conference on Multimedia Modeling (2019) 1

20. Tomar, S.: Converting video formats with ffmpeg. Linux Journal (2006) 5

21. Tunai, P.M., Alexandra, B.A., Maia, H.: A contrast-guided approach for the enhancement of low-lighting underwater images. Journal of Imaging (2019) 3

22. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1

23. Youngjae, Y., Jongseok, K., Gunhee, K.: A joint sequence fusion model for video question answering and retrieval. In: Proceeding of European Conference on Computer Vision (ECCV) (2018) 1

24. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: AAAI Conference on Artificial Intelligence (2018) 1

25. Zhuang, P., Wang, Y., Qiao, Y.: Wildfish: A large benchmark for fish recognition in the wild. In: Proceeding of ACM Multimedia Conference on Multimedia Conference (2018) 3