VIRET tool keyword search at TRECVID 2019 AVS task

Jakub Lokoč, Tomáš Souček, František Mejzlík, Ladislav Peška
SIRET Research Group, Department of Software Engineering
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

ABSTRACT

This paper presents details of a frame-based keyword search component that was used for TRECVID 2019 Ad-hoc Video Search with manually-assisted querying. The component is part of the VIRET framework primarily developed for interactive known-item search. For each task one query was formulated and the query was used to form four different runs by applying different scoring functions.

1 INTRODUCTION

VIRET [7] is a frame based known-item search interactive video retrieval framework that enables users to formulate multi-modal temporal queries and interactively browse scenes in the result set. Whereas VIRET is designed to participate at the interactive setting of the Video Browser Showdown (VBS) [5, 6] and Lifelog Search Challenge (LSC) [4] events, some of the system components can be tested also at the TRECVID evaluation campaign [1, 2]. Hence, we joined TRECVID 2019 to obtain a comparison with state-of-the-art methods for Ad-hoc Video Search (AVS) tasks that could be potentially used also for known-item search initialization with textual queries at VBS and LSC challenges.

In this paper, we detail the keyword search approach that was used to obtain ranked result sets for manually assisted AVS tasks. We would like to emphasize that the VIRET framework currently uses just a limited set of labels from a deep classification network and do not specialize on automatic action recognition. Hence, in many cases the manually assisted query reformulation was highly limited with the set of supported labels. The provided queries consisted merely from sets of selected labels that were close to the searched topic.

2 AUTOMATIC ANNOTATION

Currently, VIRET framework employs the NasNet large network [10] that was retrained using an own set of basic 1243 classes/labels $l \in C$ manually selected from ImageNet [3] and Places dataset [9]. The class selection from ImageNet involved an automatic grouping of a set of candidate classes based on their visual similarity (employing a mean deep vector representing each class [7]). In addition, a set of WordNet[1] hypernyms was used to extend the set of labels.

Given a retrained network, the automatic annotation process assigns the output vector $x \in \mathbb{R}^{1243}$ to each selected frame (VIRET operates on the set of frames selected during the data pre-processing phase). However, to better address the multi-labeled nature of the underlying data, the softmax transformation was replaced during the extraction process by the following transformation:

$$s_l = \frac{((x_l - x_{min})/(x_{max} - x_{min}))}{\sum_{\forall m \in C}((x_m - x_{min})/(x_{max} - x_{min}))}, \tag{1}$$

where x_l corresponds to the network output (before softmax) for the basic class label l and x_{min}, x_{max} represent minimal and maximal values in dimensions of x. In our experiments focusing on known-item search [8], this form of normalization outperformed softmax for queries comprising multiple labels.

3 RELEVANCE SCORING

Since a basic class label could be also a hypernym of another basic class label, we will unify the notation and denote every provided query label h as a set $H = \{\bar{l} \in C : h \text{ is hypernym of } \bar{l}\}$. For a set Q of query labels, we considered the following two aggregations to obtain score of each selected frame f_i:

$$score^{f_i}_{mult}(Q) = \prod_{\forall H \in Q} \sum_{\forall l \in H} s^{f_i}_l \tag{2}$$

$$score^{f_i}_{sum}(Q) = \sum_{\forall H \in Q} \max_{\forall l \in H} s^{f_i}_l \tag{3}$$

The computed relevance scores were used to sort selected frames and based on this ordering, official temporal segments were detected for runs C_A_SIRET.19_1 (equation 2) and C_A_SIRET.19_3 (equation 3). In addition, we tested whether temporal queries[2] [7] could improve the effectiveness of the search. Hence, in runs C_A_SIRET.19_2 (equation 2) and C_A_SIRET.19_4 (equation 3), we formulated a temporal keyword query by repeating the query ($Q_1 = Q_2$) for a following selected frame (maximally k-th). Whereas temporal queries are usually used in known-item search to target two different frames/shots, in this case the motivation was to try temporal queries to target consecutive selected frames with the same semantic content.

4 DISCUSSION

The results show several observations:

- Our simple keyword search approach was outperformed by current state-of-the-art methods focusing on Ad-hoc search. Especially in tasks involving actions performed by one or multiple persons, where our limited set of supported labels is not sufficient to distinguish between different scenes or for example colors of objects. On the other hand, there were several tasks where our approach performed well. For example, in task 618 “Find shots of coral reef underwater” all four runs found more than 471 inferred hits at depth 1000. From

[1] https://wordnet.princeton.edu/

[2] Temporal query consists of two query sets Q_1 and Q_2, where each $score^{f_i}$ for query Q_1 is combined with $max(score^{f_{i+1}}, \ldots, score^{f_{i+k}})$ for query Q_2.

TRECVID '19, 2019, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

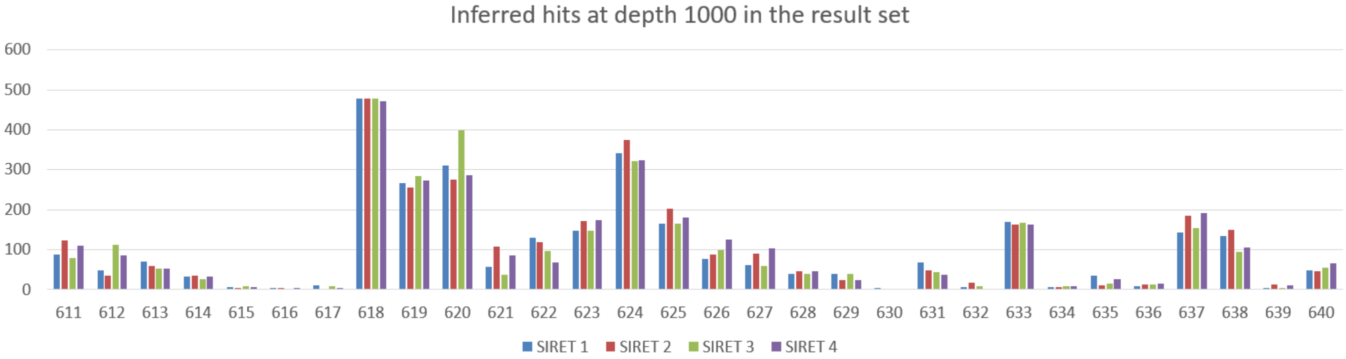

Figure 1: Results at manually assisted Ad-hoc search tasks.

this point of view, our method demonstrates what results can be obtained with a simple combination of related labels.

- There were no large differences in inferred hits at depth 1000 between the four tested relevance scoring approaches (see Figure 1). In some cases (e.g., tasks 611, 621, or 627), the temporal queries helped to improve the number of inferred hits, while in some cases the temporal queries decreased the number (e.g., 612, 619, or 622). The performance of both relevance scoring approaches (equation 2 and 3) is also mostly similar, except a few cases (e.g., tasks 612, 620, or 638). Let us note that these observations sometimes differ for hits at depth 100.

To sum up the results, the keyword search approach currently used by VIRET can be often employed to find a few examples of the searched topic even with just manually assisted settings (i.e., without observing results during query formulation), however, the recall is mostly not high in comparison to state-of-the-art Ad-hoc search systems participating at TRECVID. Let us also note that at the Video Browser Showdown AVS evaluation, the performance of the VIRET prototypes was competitive with respect to other VBS systems in an interactive setting [6], where users can reformulate the query based on observed results, employ additional retrieval models and check temporal context of returned scenes to collect relevant shots. Anyway, according to this comparative evaluation, we plan to investigate state-of-the-art Ad-hoc video search approaches and evaluate their effectiveness on known-item search tasks.

ACKNOWLEDGMENTS

This paper has been supported by Czech Science Foundation (GAČR) project 19-22071Y and by Charles University grant SVV-260451.

REFERENCES

[1] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, and Georges Quénot. 2019. TRECVID 2019: An evaluation campaign to benchmark Video Activity Detection, Video Captioning and Matching, and Video Search & retrieval. In *Proceedings of TRECVID 2019*. NIST, USA.

[2] George Awad, Asad Butt, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Quénot, Maria Eskevich, Roeland Ordelman, Gareth J. F. Jones, and Benoit Huet. 2017. TRECVID 2017: Evaluating Ad-hoc and Instance Video Search, Events Detection, Video Captioning and Hyperlinking. In *Proceedings of TRECVID 2017*. NIST, USA.

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[4] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Duc-Tien Dang-Nguyen, Michael Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoč, and Wolfgang Hürst. 2019. [Invited papers] Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7, 2 (2019), 46–59. https://doi.org/10.3169/mta.7.46

[5] Jakub Lokoč, Werner Bailer, Klaus Schoeffmann, Bernd Münzer, and George Awad. 2018. On Influential Trends in Interactive Video Retrieval: Video Browser Showdown 2015-2017. *IEEE Trans. Multimedia* 20, 12 (2018), 3361–3376. https://doi.org/10.1109/TMM.2018.2830110

[6] Jakub Lokoč, Gregor Kovalčík, Bernd Münzer, Klaus Schöffmann, Werner Bailer, Ralph Gasser, Stefanos Vrochidis, Phuong Anh Nguyen, Sitapa Rujikietgumjorn, and Kai Uwe Barthel. 2019. Interactive Search or Sequential Browsing? A Detailed Analysis of the Video Browser Showdown 2018. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1, Article 29 (Feb. 2019), 18 pages. https://doi.org/10.1145/3295663

[7] Jakub Lokoč, Gregor Kovalčík, Tomáš Souček, Jaroslav Moravec, and Přemysl Čech. 2019. A Framework for Effective Known-item Search in Video. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. ACM, New York, NY, USA, 1777–1785. https://doi.org/10.1145/3343031.3351046

[8] Jakub Lokoč, František Mejzlík, Tomáš Souček, and Ladislav Peška. [n. d.]. Evaluating keyword-based known-item search models based on deep classification networks.

[9] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[10] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2017. Learning Transferable Architectures for Scalable Image Recognition. *CoRR* abs/1707.07012 (2017). arXiv:1707.07012 http://arxiv.org/abs/1707.07012