Enhanced VIREO KIS at VBS 2018

Phuong Anh Nguyen(✉), Yi-Jie Lu, Hao Zhang,
and Chong-Wah Ngo

City University of Hong Kong, Kowloon, Hong Kong
{panguyen2-c,yijie.lu,hzhang57-c}@my.cityu.edu.hk,
cscwngo@cityu.edu.hk

Abstract. The VIREO Known-Item Search (KIS) system has joined the Video Browser Showdown (VBS) [1] evaluation benchmark for the first time in year 2017. With experiences learned, the second version of VIREO KIS is presented in this paper. Considering the color-sketch based retrieval, we propose a simple grid-based approach for color query. This method allows the aggregation of color distributions in video frames into a shot representation, and generates the pre-computed rank list for all available queries which reduces computational resources and favors a recommendation module. With focusing on concept based retrieval, we modify our multimedia event detection system at TRECVID 2015 in VIREO KIS 2017. In this year, the concept bank of VIREO KIS has been upgraded to 14K concepts. An adaptive concept selection, combination and expansion mechanism, which assists the user in picking the right concepts and logically combining concepts to form more expressive query, has been developed. In addition, metadata is included for textual query and some interface designs are also revised for providing a flexible view of results to the user.

Keywords: Video search · Known-Item Search · Color sketch query
Concept query · Concept selection · Concept combination

1 Introduction

In VBS 2017, the first version of VIREO KIS system [2] has achieved the 4th rank over 6 participants on Known-Item Search task and got the best result on Adhoc Video Search (AVS) task. This system includes three modalities: query by color sketch, by edge sketch, by concept that mainly relies on the detection of audio-visual objects. Based on the lesson learnt from VBS 2017, we have investigated the limitations of our system and proposed solutions in this paper.

First, query by sketch is relied on the color signatures feature which was originally proposed and implemented by SIRET team [3, 4]. The main idea is extraction of color circles from video frames for matching with the color circles sketched by the user using a spatial grid indexing technique. The effectiveness of color signatures feature depends on the temporal sampling rate for feature extraction: dense sampling brings accurate results with the expense of low efficiency due to dramatic increase of candidates for matching; master shot key-frames sampling processes less number of frames for color circles extraction which reduces retrieval accuracy. To deal with this problem, we propose a new method to capture the color information of video frames and enable

K. Schoeffmann et al. (Eds.): MMM 2018, Part II, LNCS 10705, pp. 407–412, 2018.
https://doi.org/10.1007/978-3-319-73600-6_42

query by color sketch. In this approach, a uniform grid is placed over a video frame and the color distribution of each cell is calculated. For retrieval, the user needs to formulate a cell-based query including both cell's location and dominant color for matching. Using this method, the color distribution of video frames can be aggregated into shot-based representation which significantly reduces the matching expense comparing to frame-based representation. Moreover, a set of colors available for query is defined enabling pre-computing the matching result for every single cell-based query. The pre-computed result then favors a recommendation module which provides to the user the color distribution information of the dataset. This recommendation module has been inspired by our previous work on a simulation framework for color sketch [5].

Second, concept selection is one of the critical part of the concept based KIS system. Previously, our system relied on an automatic concept selection module combined with the user's evaluation for concept re-selection and concept-weight estimation. In practice, the concept selection can be done on the fly with a simple recommendation module which directly maps the user input to the concept bank. To enrich the query, we follow the Waseda team's report in AVS task 2016 [6] and provide the AND, OR, NOT combination of concepts. Also, a recommendation module has been developed to provide a list of co-occurrence concepts for filtering. We also refer to an existing tags tokenizing interface in website development to implement the user interface specifically for the concept selection.

Third, to improve the searchability of textual query, we include all metadata including video name, video description, speech and on-screen text for retrieval. Also, a compact video key-frame representation is also developed using the idea of video thumbnail preview.

2 Simplified Color Sketch Query

As aforementioned, the color signatures feature provides a flexible mechanism for defining a query with arbitrary position and color. Because the user can freely put color circles in any position with any color, the matching calculation must be done on the fly which requires computing and sorting distances from the query color circles to color circles in the database. This approach leads to a tedious matching calculation when the size of the dataset grows larger. Based on a study from Johns Hopkin University which shows that the user's memories for colors are biased in favor of his/her "best" versions of basic colors [7], we combine with the grid-based color sketch in [8] to define a fixed set of queries based on the basic colors. Relying on this set of queries, the user can construct an approximate version of their visual memory as an input to the video retrieval system. The advantage of using this set of queries is that the matching phase can be done in advance for all pre-defined queries. Hence, it creates more room to calculate result for the combination of multiple queries and enables recommendation to the user.

At the first step, we select a fixed set of colors $C = \{c_1, \ldots, c_n\}$ as the available query colors. A uniform grid is placed over the color sketch area generating a list of cells $P = \{p_1, \ldots, p_m\}$. To construct a query q, the user needs to specify both the cell position p and the color c of that cell, i.e. $q = (p, c)$.

In the matching phase, the distance from a query q to a video frame f is calculated by taking the average of the pixel-wise Euclidean distances between the query color q_c and the color at the same query cell p in frame f.

$$D(q, f) = \frac{1}{n_p}\sum_{x=1}^{n_p} d(c_x, q_c) \tag{1}$$

where c_x is the color of pixel x in cell p of f and n_p is the number of pixels in cell p.

As a query's cell position and color have been fixed, we have totally $Q = \{q_1, \ldots, q_k\}$ available queries with $k = n * m$, where n is the number of selected colors and m is the number of cells generated by the uniform gird. Using (1), the distances from each query in Q to all video frames in the database can be calculated in advance and indexed.

Based on (1), we aggregate the frames in the same shot to generate the distances from each query in Q to all shots. The aggregation can be done by taking the average or the minimum of the distances over all frames in a shot $s = \{f_1, \ldots, f_h\}$. In details,

$$DS(q, s) = \frac{1}{h}\sum_{i=1}^{h} D(q, f_i) \tag{2}$$

$$\text{or} \quad DS(q, s) = min_{i=1}^{h} D(q, f_i) \tag{3}$$

where $D(q, f_i)$ is the distance from a query q to the i-th frame and h is the number of frames in shot s. The average or the minimum function can be actively selected by the user depending on his/her attention. For example, if the user focuses in the areas where the color does not change in the shot, the average function will be the best choice. In contrast, if the user focuses on the continuously changing color in the shot, the minimum function should be used. As a result, we have pre-computed distances to all video shots from all queries in Q using both (2) and (3) approaches.

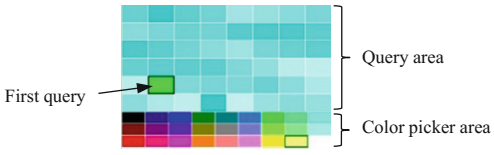

Fig. 1. Dynamic update of color query area for system recommendation. In this example, the user picks yellow color as the second color after specifying the grid position of green color. The color area is updated with each cell indicating the available number of shots to be retrieved. The lighter color means a more unique cell with less number of candidate shots to be browsed. (Color figure online)

Using these pre-computed distances, the result can be quickly retrieved as well as recommendation can be made. Assuming that the user has already put a query q and intends to put a second query q', we require the user to specify the cell's color c' first, followed by the location p' of q'. Then, the pre-computed distances for all combinations of $q' = (p', c')$ can be loaded and combined with the existing distances from q. Finally, a distance threshold is defined which can help counting the number of shots that lie in a

specific range from those queries. These numbers provide the user the ideas about the distribution of the result that he/she will get in the next query.

According to our interface design for color sketch (see Fig. 1), the user needs to select a color from the color picker area first and then click on the query area to place the query, i.e. for every query q', the user will specify the color c' as the input to the recommendation module. The distribution of the result that the user gets from q' is shown in all non-placed query cells in the query area indicated by the intensity of each cell. The darkness of a query cell is calculated by normalizing the number of shots in defined distance threshold by the maximal value of these numbers over all query cells.

3 Enhanced Concept Selection, Combination and Query Expansion

To improve the effectiveness of the concept based modality, we first replace the previous concept bank [9] by the larger one with 14,046 concepts including: ImageNet with 12,988 concepts, MIT Places with 205 concepts [10], TRECVID SIN Task with 346 concepts [11] and Research collection with 497 concepts [12]. With a large number of concepts, the process of selecting the query concepts becomes more important. In the previous study, the retrieval accuracy is improved by the user's engagement in the concept selection process. With the advantages of an interactive system, the performance of concept selection can be significantly improved by letting the user pick the right concepts from the start. Then, with the understanding of the co-occurrence concepts from the retrieved results, the user can refine the query by adding more specific concepts or removing un-related concepts. To accomplish these two tasks, we implement the system as follows:

Concept selection. The target of this step is to force the user to directly select concepts from the concept bank. This part has been done by two steps: when the user inputs the text, the system directly show a list of concepts which shows the concepts start with the input text using a drop-down list; when the user finishes typing but the concept does not exist in the concept bank, a list of nearest concepts in WordNet is updated to the drop-down list for use selection.

Concept combination. This function allows the user to define a specific query based on three logic combinations: AND, OR, NOT. The proposed method of Waseda team [6] for these logic combinations has been implemented in VIREO KIS. In order to provide a user-friendly interface, an interface for multi-tags selection which is commonly used by search engines, is developed to help the user formulate query easily. Instead of treating every concept as a tag, we consider each logic combination of concepts as a tag, each tag comes with a background color which shows the meaning of the combination. All the stand-alone tags are combined using the AND combination. The user can easily add more concepts combination or remove existing combination in a tokenizing area (see Fig. 2).

Concept query expansion. After feeding the query into the retrieval systems, the rank list of video shots is retrieved together with a list of related concepts which largely

impact the result. The related concepts are generated by evaluating the co-occurrence concepts [13] based on their influence and frequency in the top-k shots retrieved. From these concepts, the user can pick more concepts and update query. We expect that the user will mostly use the AND function to create more detail query and the NOT function to filter out unrelated results.

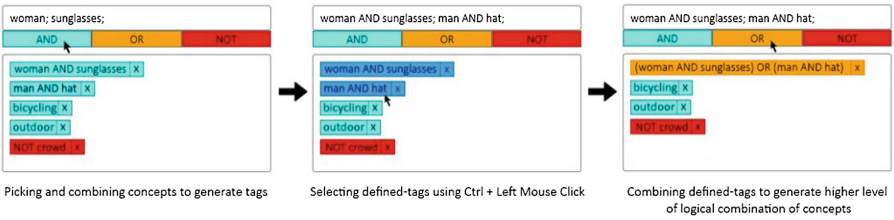

Fig. 2. The combination of concepts. After picking a list of concepts, the user can choose the logical combination by clicking on the corresponding button. Each tag comes along with a removal button ("X") such that the user can modify the query interactively.

4 Metadata for Textual Query and Interface Design

Metadata. From VBS 2017, we figured out that the metadata is very helpful in many cases. Hence, we deploy a text-based retrieval module using Lucence to index the text extracted from file name, video description, automation speech provided with the video data and the optical character recognized by Tesseract OCR Engine [14].

Interface design. The existing interface is based on the design of SIRET team [3, 4] with each video being shown as a row accompanied with its temporal context. The display occupies a lot of screen space resulting in only 10 videos being shown on screen. This way of presentation reduces the effectiveness of the browsing phase in the early stages of retrieval for large-scale dataset, so we go back to the traditional representation which shows the candidate master-shot key-frames only. To provide a quick judgement to the user, we let each master-shot key-frame become a dynamical image where frames in shot are shown continuously on mouse hover. We also employ a preview panel which allows the user to view the summarization of a video using master shot key-frames as well as preview the video content with fast forward speed.

5 Conclusion

The 2018 version of VIREO KIS has significant improvement in term of resource consumption and query latency comparing to the last year version. With a simplified color sketch query with recommendation module, the user now has more understanding on the dataset as well as has more guidance in formulating queries. Using a large concept bank with a flexible concept selection, combination and expansion mechanism provides an interesting playground for the user in adapting queries especially in dealing with large-scale dataset. And finally, metadata brings more options in revising queries as well as the new browsing interface is more compact for providing a broader view on retrieval results.

Acknowledgment. The work described in this paper was supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11210514, 11250716).

References

1. Cobârzan, C., Schoeffmann, K., Bailer, W., Hürst, W., Blažek, A., Lokoč, J., Vrochidis, S., Barthel, K.U., Rossetto, L.: Interactive video search tools: a detailed analysis of the video browser showdown 2015. Multimed. Tools Appl. **76**(4), 5539–5571 (2017)
2. Lu, Y.-J., Nguyen, P.A., Zhang, H., Ngo, C.-W.: Concept-based interactive search system. In: Amsaleg, L., Guðmundsson, G.Þ., Gurrin, C., Jónsson, B.Þ., Satoh, S. (eds.) MMM 2017. LNCS, vol. 10133, pp. 463–468. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51814-5_42
3. Lokoč, J., Blažek, A., Skopal, T.: Signature-based video browser. In: Gurrin, C., Hopfgartner, F., Hurst, W., Johansen, H., Lee, H., O'Connor, N. (eds.) MMM 2014. LNCS, vol. 8326, pp. 415–418. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-04117-9_49
4. Blažek, A., Lokoč, J., Matzner, F., Skopal, T.: Enhanced signature-based video browser. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015. LNCS, vol. 8936, pp. 243–248. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-14442-9_22
5. Lokoč, J., Phuong, A.N., Vomlelová, M., Ngo, C.-W.: Color-sketch simulator: a guide for color-based visual known-item search. In: Cong, G., Peng, W.-C., Zhang, W.E., Li, C., Sun, A. (eds.) ADMA 2017. LNCS, vol. 10604, pp. 754–763. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69179-4_53
6. Ueki, K., Kikuchi, K., Saito, S., Kobayashi, T.: Waseda at TRECVID 2016: ad-hoc video search. In: TRECVID 2016 Workshop, Gaithersburg, MD, USA (2016)
7. Bae, G.Y., Olkkonen, M., Allred, S.R., Flombaum, J.I.: Why some colors appear more memorable than others: a model combining categories and particulars in color working memory. J. Exp. Psychol. Gen. **144**(4), 744–763 (2015)
8. Wang, J., Hua, X.-S.: Interactive image search by color map. ACM Trans. Intell. Syst. Technol. **3**(1), Article Id 12 (2011)
9. Lu, Y.J., Zhang, H., de Boer, M., Ngo, C.W.: Event detection with zero example: select the right and suppress the wrong concepts. In: ACM ICMR (2016)
10. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems, pp. 487–495 (2014)
11. Zhang, W., Zhang, H., Yao, T., Lu, Y., Chen, J., Ngo, C.-W.: VIREO @ TRECVID 2014: instance search and semantic indexing. In: NIST TRECVID Workshop (2014)
12. Strassel, S., Morris, A., Fiscus, J., Caruso, C., Lee, H., Over, P., Fiumara, J., Shaw, B., Antonishek, B., Michel, M.: Creating HAVIC: heterogeneous audio-visual internet collection. In: Chair, N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) LREC, Istanbul, Turkey, May 2012. ELRA (2012)
13. Sigurbjornsson, B., Zwol, R.V.: Flickr tag recommendation based on collective knowledge. In: Proceeding of ACM Intelligent World Wide Web Conference, pp. 327–336 (2008)
14. Smith, R.: An overview of the tesseract OCR engine. In: Proceeding of 9th International Conference on Document Analysis & Recognition (2007)