VIREO @ Video Browser Showdown 2019

Phuong Anh Nguyen[1], Chong-Wah Ngo[1], Danny Francis[2], and Benoit Huet[2]

[1] Computer Science Department, City University of Hong Kong, China
panguyen2-c@my.cityu.edu.hk
cscwngo@cityu.edu.hk
[2] Data Science Department, EURECOM, France
danny.francis@eurecom.fr
benoit.huet@eurecom.fr

Abstract. In this paper, the VIREO team video retrieval tool is described in details. As learned from Video Browser Showdown (VBS) 2018, the visualization of video frames is a critical need to improve the browsing effectiveness. Based on this observation, a hierarchical structure that represents the video frame clusters has been built automatically using k-means and self-organizing-map and used for visualization. Also, the relevance feedback module which relies on real-time support-vector-machine classification becomes unfeasible with the large dataset provided in VBS 2019 and has been replaced by a browsing module with pre-calculated nearest neighbors. The preliminary user study results on IACC.3 dataset show that these modules are able to improve the retrieval accuracy and efficiency in real-time video search system.

Keywords: video visualization, video retrieval, video browser showdown

1 Introduction

The VIREO team has participated in the Video Browser Showdown [1] for two recent years and finished with the best ad-hoc tasks in 2017 [2] and the best visual known-item search tasks in 2018 [3]. The tool of the VIREO team provides two searching methods: color-sketch-based and text-based. Using these two functions, users are able to input queries based on their understanding and their memory. Getting the initial results of queries, users will repeatedly update queries and judge results in a loop until finding correct answers.

In 2018, the VIREO team ranked at 6^{th} position over 9 teams participating in ad-hoc tasks although the concept bank used for text-based retrieval is rich with 14K concepts. Digging down the problem, we concluded that the tool was lacking of the ability allowing users to explore the dataset with alignments. This is a well-known research topic named image visualization.

From the starting point of Video Browser Showdown, many teams integrated the visualization module to the video retrieval tool and achieved remarkable results. KLU - AAU video browser is the winner in 2012 and the tool relies on the ability of humans to track several video streams by hierarchical or parallel

browsing methods [4]. The following version used a 3D thumbnail ring arranged by color similarity for video browsing [5]. The SIRET team is the winner for two consecutive years in 2014 and 2015. In the version of 2015 [6], the team focused on browsing by using the detected shot boundaries and key-frames displayed in a compact way. The HTW team is the winner in 2016 [7]. The team used a novel browsing approach based on graph hierarchy and visually stored image maps. Visual and semantic features learned from convolutional neural network (CNN) is used to predefine graph structure and perform clustering. Then, the graph is projected to 2D plane, performs discrete optimization and generates hierarchies. These results show that the visualization technique for browsing is a critical module for the video retrieval tool.

Following aforementioned approaches, we propose a simple way to construct the hierarchical visualization of the dataset. Basically, the two main factors determining the effectiveness of browsing using the visualization module are the features and the clustering algorithms used for constructing the hierarchy. In proposed approaches, CNN features and a color histogram are used because their robustness has been proven in image retrieval task. For clustering, k-means and self-organizing-map (SOM) [8] are considered. Setup details and preliminary experiment results are described in the next section.

2 Data visualization for browsing

2.1 Feature extraction

From the video key-frames, we extract two types of features including CNN feature from Deep Residual Net (ResNet) [9] and the color histogram in RGB color space. More precisely, we use ResNet50 which is a 50 layers Residual Network, take the pool 5 layer feature map then perform PCA to reduce the dimension to 128 for clustering. As the color histogram can be built for any kind of color space, the RGB color space is used for simplicity. The pixel values from each channel are discretized into 8 bins and used to form up a 3D color histogram for clustering.

2.2 Clustering and hierarchy construction

According to our experiences, the user can observe a limited number of images on the screen at a time. From that, judging all the video key-frames without any alignment is a tedious task for the user. With hierarchical partitioning, the extracted features which contain the color and semantic information are grouped and aligned in an intuitive way for browsing. This hierarchy is built once using all the dataset, while searching and judging, the user can directly navigate to the position of selected shot in the hierarchy to expand the searching area.

A grid with 8 rows and 10 columns is defined on the browsing window, each cell represents an image to the user. Each image is the center of one cluster and it accounts for all the images belonging to that cluster. The user can use

left-mouse-click on the images to go down one layer and use right-mouse-click to go up one layer in the hierarchical representation. With this design, users can quickly judge and navigate between clusters to refine browsing results.

To build up the hierarchy, we run self-organizing-map and k-means on the dataset to get 8*10 clusters and recursively do the clustering on the generated clusters with the same parameters (the number of clusters) while the number of images in the new cluster is above 80. The image which is chosen as the representation image is the one closest to the cluster center.

2.3 Preliminary experiment

For our experiments, the standard IACC.3 video dataset which was introduced in 2016 for video retrieval is used. More precisely, we use the master-shot key-frames provided by TRECVID (TREC Video Retrieval Evaluation series) which contains 335,944 key-frames. By the definition of the master-shot key-frame, key-frames are expected to provide rich information for representing the content of the video shots. Using these images, we treat the problem of video retrieval as image retrieval.

Fig. 1. The visualization of images at the top level of the hierarchy using different clustering algorithms and different types of feature.

Intuitively looking at the clustering results on the tool's interface (in Fig. 1), the clustering result of the SOM is shown in a way that makes the user un-

derstand easily with adjacent images a.k.a. clusters look similar. This is the characteristic of SOM that provides a topology preserving mapping from a high dimensional space to a 2D space. The visualization using k-means looks like chaos and creates difficulties for the user in navigation. Besides, comparing between RGB color histogram and ResNet feature for visualization using SOM, RGB color histogram visualization looks more reasonable because the color matches with human vision.

To have a more precise evaluation, a preliminary user study has been held by letting 5 users find some queries using the tool and the searching time has been used to compare and select the best setting of the visualization tool. The selected users are novice users who are trained to use the browsing tool with a ten minutes tutorial and experienced one example before participating in the real user study. In order to provide a better view of the results, two color-favoring-images and two semantic-favoring-images are randomly picked as the queries. Selected queries are shown in in Fig. 2.

Fig. 2. Selected images for the user study. Two images on the left favor color, two images on the right favor semantic features.

As shown in Table 1, using the same type of features, the visualization using SOM outperforms other visualizations in most cases thanks to its topology preserving ability. It also shows that both types of features including color histogram and ResNet features are useful in visualization for searching. The shortest time to find color-favoring-images belongs to the RGB color histogram, the shortest time to find semantic-favoring-images belongs to the ResNet feature. These results lead us to the decision of using SOM with color histogram and ResNet feature for the visualization tool.

3 Video retrieval tool description

Besides integrating the visualization module into the video retrieval tool, the searching module mostly stays the same as the tool that used in VBS 2018, including:

Color-sketch based query. This is the essential module which brings the VIREO team the best result in 2018 visual known-item-search task. With pre-calculated ranking list for all available queries which is the combination of cells

Table 1. The user study result showing the average searching time of participants who manage to find the correct answer. The number in the parenthesis shows the number of users successfully finding the query image within 3 minutes (180 seconds). The infinity symbol (∞) shows that no user manages to find the query image in allocated time.

Query	SOM+RGB	k-means+RGB	SOM+ResNet	k-means+ResNet
No. 1 (color)	**23.8s (5)**	106.2s (5)	**45.4s (5)**	127.75s (4)
No. 2 (color)	**60.8s (5)**	∞ (0)	∞ (0)	∞ (0)
No. 3 (semantic)	**98s (5)**	17.5s (2)	**18.4s (5)**	30.6s (5)
No. 4 (semantic)	**43.33s (3)**	124.5s (2)	**12.4s (5)**	23.4s (5)

on the uniform grid and the available query colors, the retrieval can be done in real-time. The combination of queries is encouraged to reduce the number of retrieved samples for judging. Two modes of queries are provided to support different attentions of the user: the color distribution of frame-based and shot-based. Because of the robustness and the expand-ability of this approach, the module is kept as the original version.

Text based query. In this module, two searching modes are provided: free text search for meta-data and exact text search for concept. The meta-data contains the video name, description, extracted speech and on-screen text. The concept search uses the 14K concept bank which provides general concepts to fine-grained concepts. However, with the striking development of object detection techniques using CNN, the object detection result of 80 common objects in context (from COCO dataset [10]) is extracted using YOLOv3 [11] and added to the exact text search function.

Filtering. Reducing the number of samples for judging can save a lot of time and favor all the searching modules. Hence, filtering module is helpful in many cases. In the tool, two basic filtering functions are provided: black-borders filter and black and white filter.

Relevance feedback. Originally, the relevance feedback module has been built based on real-time classification using SVM on the ResNet50 features. The user can pick positive and negative samples then get the classification result for judging. Usually, this module takes 4-5 seconds to generate new result on IACC.3 dataset. This is unfeasible when the dataset size dramatically increases and not suitable with the coming V3C1 dataset in VBS 2019. Also, collecting positive and negative samples in the judging process is not reasonable to the user. Instead of collecting samples, the user can directly explore the dataset to find similar samples when they look at any positive sample. This process is not time-consuming because the nearest samples for each master-shot key-frame can be calculated in advance. As a result, the relevance feedback module has been replaced by a list of top 1000 nearest neighbors of the picked sample by the user to expand the browsing space.

4 Conclusion

In the latest version, we focus on improving the effectiveness of browsing phase by proposing a simple method to construct the cluster hierarchy of video-frames. This method builds up the hierarchy based on the typical distribution of the dataset and supports the user in understanding the dataset. Besides, the replacement of the relevance feedback module is aiming to help the user expanding their searching space using any positive sample. With current development of the tool, we are looking forward to see how the system works in VBS 2019.

Acknowledgments. 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 11250716, F-CityU104/17).

References

1. Cobârzan, C., Schoeffmann, K., Bailer, W. et al: Interactive video search tools: a detailed analysis of the video browser showdown 2015. In Multimedia Tools and Application, vol. 76, pp. 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).
3. Nguyen, P.A., Lu, YJ., Zhang, H., Ngo, CW.: Enhanced VIREO KIS at VBS 2018. In: Schoeffmann K. et al. (eds) MultiMedia Modeling. MMM 2018. LNCS, vol. 10705, pp. 407-412. Springer, Cham (2018).
4. Del Fabro, M., Münzer, B., Böszörmenyi, L.: AAU Video Browser with Augmented Navigation Bars. In: Li S. et al. (eds) Advances in Multimedia Modeling. LNCS, vol. 7733, pp. 544-546. Springer, Berlin, Heidelberg (2013).
5. Schoeffmann, K., Ahlström, D., Böszörmenyi, L.: Video Browsing with a 3D Thumbnail Ring Arranged by Color Similarity. In: Schoeffmann K., Merialdo B., Hauptmann A.G., Ngo CW., Andreopoulos Y., Breiteneder C. (eds) Advances in Multimedia Modeling. MMM 2012. LNCS, vol. 7131, pp. 660-661. Springer, Berlin, Heidelberg (2012).
6. 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).
7. Barthel, K.U., Hezel, N., Mackowiak, R.: Navigating a graph of scenes for exploring large video collections. In: Tian Q., Sebe N., Qi GJ., Huet B., Hong R., Liu X. (eds) MultiMedia Modeling. MMM 2016. LNCS, vol. 9517, pp. 418-423. Springer, Cham (2016).
8. T. Kohonen: The self-organizing map. In Proceedings of the IEEE, vol. 78, no. 9, pp. 1464-1480 (1990).
9. K. He, X. Zhang, S. Ren, and J. Sun: Deep residual learning for image recognition. In: CVPR. IEEE Computer Society, pp. 770–778 (2016).
10. Lin TY. et al: Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. LNCS, vol. 8693, pp. 740-755. Springer, Cham (2014).
11. Redmon, Joseph, Farhadi, Ali: YOLOv3: An Incremental Improvement. arXiv:1804.02767 (2018).