Faceted Navigation for Browsing Large Video Collection

Zhenxing Zhang, Wei Li, Cathal Gurrin, Alan F. Smeaton

Insight Centre for Data Analytics
School of Computing, Dublin City University
Glasnevin, Co. Dublin, Ireland
{zzhang,wli,cgurrin,asmeaton}@computing.dcu.ie
https://www.insight-centre.org

Abstract. This paper presents a content-based interactive video browsing system to address the challenge in a live video search competition to find specific video clips from a large video collection under time constraints. Since the target of this evaluation forum is to evaluate and demonstrate the development of interactive video search tools, we do not need to consider if the most commonly used query-by-example or query-by-text approaches for large-scale image/video retrieval are appropriate in this scenario. In this paper, we describe an interactive video retrieval system which employs the concept filters and faceted navigation to aid users quickly and intuitively locate the interested content when browsing in a large video collections based on automatically extracted semantic concepts, object labels and attributes from video content.

Keywords: Multimedia Information Retrieval, Faceted Navigation

1 Introduction

The target of 2015 Video Browser Showdown (VBS) [1] task is to encourage the participants to design and develope the interactive search and browse systems which could help users to explore the interested video clips quickly and effectively from a large content archives. The evaluation dataset contains around 100-200 hours of video with the content comprising of various BBC TV programmes. This dataset is released in advance for participants to help them for developing and training their interactive retrieval systems and underlying content-analysis tools.

The search tasks simulate Known-Item Search (KIS) tasks which are common in the real-world and have two search categories: Visual KIS (VKIS) task, it searches for the video clips from the data collection, the length of each video clip is around 30 seconds, which are randomly selected and displayed for the user to view; Textual KIS (TKIS) task, different from the VKIS task, TKIS searches for a specific video segment according to the search topic. These search topics will be presented to all competitors on site and the visual topics will be only played once. During the competition, each team has 3 minutes to locate the known-item video clip and the performance score will be calculated based on a combination of search accuracy and search efficiency.

Especially, in order to encourage the participate researchers to develop more creative content exploring and browsing tools, the state-of-the-art search tools, such as query-by-example or query-by-text, are not allowed in this competition.

In previous years competition, the participants have proposed various interactive systems to achieve the goal of the task from different aspects. However, an overall analysis on these approaches reveals that most of them focusing on the idea of searching/filtering and browsing the results, instead of considering of using the high-level semantic concepts or low-level visual descriptors. Based on the wide range of approaches used by previous participants [2] in the competition, we made an observation that, the subjective and imprecise user expressions pose challenges for teams to provide appropriate tools to aid users to form suitable queries in order to make use of the advanced visual retrieval capabilities which have been embedded in these interactive systems. Only the SIRET team [3] gained success for visual retrieval tasks by making good use of position-color feature signatures and designed a special color picker interface to allow users to draw coloured regions when searching for the interested video clip.

Based on previous participants work [2], in this competition, we'll focus on visual content exploration and navigation technologies, and aimed to build an interactive user interface with the abilities of browsing large sets of video content quickly and effectively. To achieve this goal, we developed an interactive exploratory search system with faceted navigation [4] based on the outputs of visual content analysis, such as automatically extract semantic concepts, object labels, and object attributes from image/video content using various approaches. To fulfill the requirement of finding the specific video segments, our system allows users to locate the video clip from multiple facets using different combinations of semantic concepts in various order.

2 Video Segmentation and Representation

To reduce computing complexity and remove the near-duplicate frames, a common method used in content-based video analysis is to segment a video into constituent shots, each of which is a temporally sequential set of video frames, with an appropriate keyframe to represent the content of each shot. In this work, we segmented the video into shots based on the shot boundaries detection technologies outlined by Pickering et al.[5]. Then we extracted the most central frame to provide a quick overview over different scenes through the large video collection, this strategy has been approved well performance in the TRECVid instance search task [6]. The average duration of a shot in our system is around 10 seconds which means that a 100 hours video collection would be abstract to about 36,000 shots. In addition, we also extracted dense frames in order to provide a detailed context for users to inspect the video segments whenever necessary.

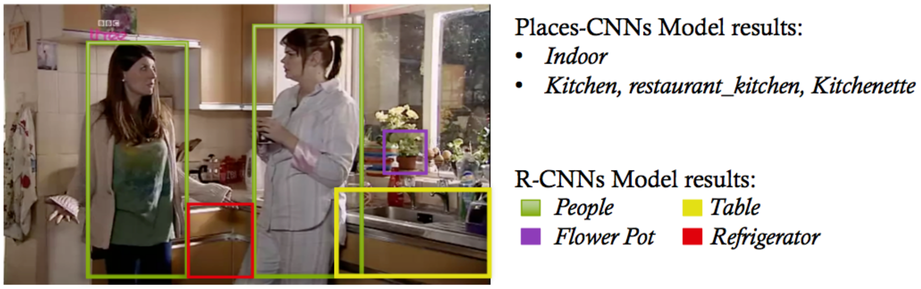

Fig. 1. An example results from visual content analysis using DCNs models. The Place-CNNs model labeled the image with overall background information, and the R-CNN model focused on label image with object categories.

3 Visual Content Analysis

The recent development in image classification and object recognition using the Deep Convolutional Networks (DCNs) has been shown to perform very well on various evaluation data sets. The learning framework, Caffe [11], along with learned models is available to use as open source, which encourages researchers to use and contribute to the framework. In our work, we employ DCNs to extract meaningful information, such as semantic concepts, object labels and object attributes, to describe the visual content of video shots. More specifically, we chose two pre-trained models to cover the wide range of possible topics in the BBC programming videos, these two models are:

R-CNN ILSVRC-2013 [7] this model includes 200 object categories, such as Person, Dog, Chair and so on. We chose this model to describe desired object information that could be easily captured by users while interpreting the search topics.

Places-CNNs [9] The second model introduced 205 scene categories to describe the overall background context for any given images, such as office, restaurant, valley, desert and so on. We chose this model to allow users to navigate the video content with straightforward environments information.

We used the Caffe Library[1] running on a server machine equipped with GeForce GTX 970 Graphic card and 16 GB RAM to do the heavy visual processing tasks. Thanks to the efficient algorithm implementation and GPU computational power, it took about two seconds to extract the visual information and label a frame into a set of text words.

Figure 1 illuminates an example result after applying the semantic visual analysis process for one random frame from the video collection. When using the R-CNNs model, the bottom up region proposals approach was employed so that we could localize and label the objects from complex scene frame. For scene recognition, the frame image was used as one input for the Places-CNNs model to label the most likely scene categories from various informative regions.

[1] http://caffe.berkeleyvision.org/

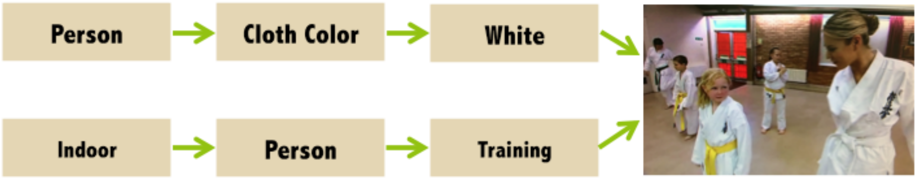

Fig. 2. This example demonstrates two possible routes to find the target video clip with text query: *A group of mostly kids practicing Karate moves indoors (in white clothes), including close-ups of a blond young woman talking to a girl, and shots showing the instructor, a bald man with glasses.*

4 Faceted Navigation

Faceted navigation is implemented in this work to support exploratory search and navigation, while aiding users to find the interested video clips quickly and effectively. Essentially faceted navigation reflects the fact that users may seek information in a number of different ways from their own understanding of the query topics. In our standard faceted navigation system, the facets are built from different object labels or scene categories which are generated during the video content analysis process. These facets in turn contain attributes by which the list of the same category can be further filtered. After the general understanding of example text query topic in the first step, users can very easily to locate the target video clips by following the provided various facets.

Compared to the previous proposed tools for this task, the main advantages of our faceted navigation system could be summarised into the following three points:

- It does not require users to manually input search query to match the high-level semantic concepts or low-level feature descriptors. This frees users from manually interpret the complex visual or text topics.
- It could take the full advantage of the advanced visual analysis approaches. Since the facets was organized and presented as filters for users to narrow down the range of search results, it could benefits from more precise and detailed content descriptions.
- Finally, it provides multiple navigation routes to help users to identify the same video clip.

Our faceted navigation relies on an underlying infrastructure that enables associations among elements of multiple types and allow users to drill down through categories and attributes naturally. The taxonomy structure of our system is constructed from the visual content of key frames and specifically addressed on three aspects: Similarity clusters, Color, and Semantic labels. To explore the content of video collection, users could choose different facets and allow the system to narrow down the candidate collection for inspection. Figure 2 demonstrates an example of using our faceted navigation to find the interested video clips from two different routes.

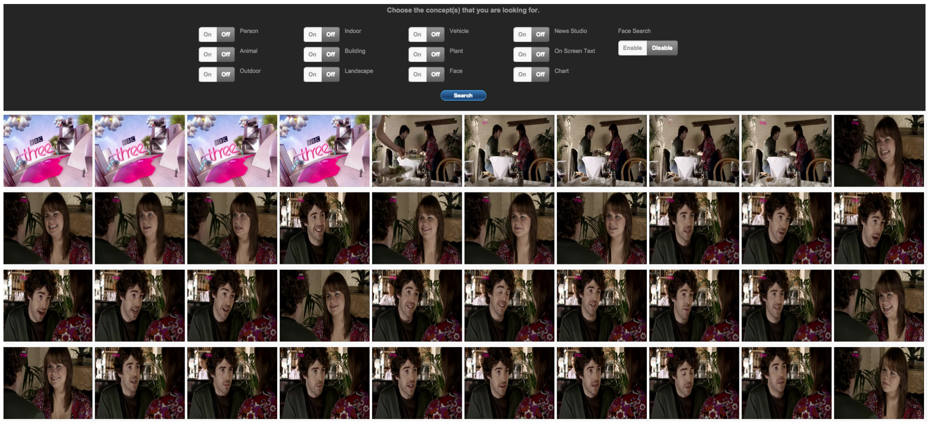

Fig. 3. Screenshot for interactive browser interface. Concept filters are provided on the top panel, dense frames are displayed in the main window for interpret video clip.

5 Browsing Interface

Based on our experience of previous participation [2], we followed two main guidelines to design the interactive user interface for this system. Firstly, the interface should be simple and straightforward to allow users efficiently browse when dealing with a large amount of scenes. Secondly, the interface should allow users to quickly inspect of the video clips by providing an overview of the context, such as the previous and following shots.

Figure 3 illustrates the initial design of our interactive interface for faceted navigation system. The user interface allows the users to trigger the search process by selecting the interested concept filters, and navigate them to explore the content by continuously ask them to contribute input to further narrow down the search range by suggesting useful facets.

6 Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289.

References

1. Klaus Schoeffmann, A User-Centric Media Retrieval Competition: The Video Browser Showdown 2012-2014, in IEEE Multimedia, Vol. 21, No. 4, Oct.-Dec., 2014, pp. 8-13 (2014)
2. Z. Zhang, R. Albatal, C. Gurrin, A. F. Smeaton. : Interactive Known-Item Search Using Semantic Textual and Colour Modalities. In: The Processing of Multimedia Modeling Conference. LNCS, vol. 8936, pp. 282-286. Springer International Publishing (2015)

3. Lokoč, J., Blažek, A., Skopal, T.: Enhanced Signature-Based Video Browser. In: The Processing of Multimedia Modeling Conference. LNCS, vol. 8936, pp. 243-248. Springer International Publishing (2015)
4. Hearst, M.A. : UIs for Faceted Navigation: Recent Advances and Remaining Open Problems. In: the Workshop on Computer Interaction and Information Retrieval, HCIR 2008. (2008)
5. Pickering, M.J., Ruger, S.M.: Evaluation of key frame-based retrieval techniques for video. Computer Vision and Image Understanding 92(2-3), 217235 (2003)
6. Z. Zhang, R. Albatal, C. Gurrin, A. F. Smeaton. : TRECVid 2013 Experiments at Dublin city University In: TREC Video Retrieval Evaluation.(2013)
7. R. Girshick, J. Donahue, T. Darrell, J. Malik. : Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
8. S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue and T. Darrell.: Open-vocabulary Object Retrieval. In: Robotics: Science and Systems (2014)
9. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva.: Learning Deep Features for Scene Recognition using Places Database. in: Advances in Neural Information Processing Systems 27 (2014)
10. G. Patterson, C. Xu, H. Su, J. Hays. : The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. In: International Journal of Computer Vision, Volume 108, Issue 1, pp 59-81. (2014)
11. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, Trevor Darrell: Caffe: Convolutional Architecture for Fast Feature Embedding. In: ACM MM Open Source Competition (2014)