Interactive Known-Item Search using semantic textual and colour modalities

Zhenxing Zhang, Rami Albatal, Cathal Gurrin, and Alan F.Smeaton

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

Abstract. In this paper, we propose an interactive video browser tool for our participation in the fourth video search showcase event. Learning from previous experience, this year we focused on building an advanced interactive interface which allows users to quickly generate and combine different styles of query to find relevant video segments. The system offers the user a comprehensive search interface which has as key features: keyword search, color-region search and human face filtering.

Keywords: Multimedia Indexing, Deep Learning, Human Face Detection, Interactive Interface

1 Introduction

Two different search categories have been proposed in this year's Video Browser Showcase: Visual Known-Item Search (KIS) and Descriptive Known-Item Search. In order to address these advanced evaluation procedures, the classic video retrieval approaches based on query images (or sketch images) may not necessarily be sufficient. Hence, we decided to improve our video retrieval system from last year's participation [1]; in this work we propose a more comprehensive, effective and efficient video browser tool that contains three main retrieval modalities: *a*) Keyword-based Retrieval, *b*) Color-based Retrieval, *c*) Human Face Filtering. By combining multiple retrieval modalities with a flexible and interactive user interface, the system enhances the user's ability to quickly locate the required video segments.

The rest of the paper is organized as follows. Section 2 discusses the different proposed retrieval modalities; then in Section 3 we describe our retrieval interface under development; finally section 4 concludes the work.

2 Retrieval modalities

In order to address the two search categories of this year's challenge, two dimensions are considered when analysing the video collection: a visual and a semantic

dimension. Hence we provide multiple retrieval modalities that explore the content of the collection from both visual and/or from semantic point-of-views. By adopting state-of-the-art content-based image retrieval technologies, users can employ one or more of the following available retrieval modalities.

2.1 Keyword-based Retrieval

A keyword-to-image approach is developed to help users in formulating textual queries by proposing keywords from a pre-defined vocabulary. The suggested keywords correspond to pre-trained classifiers used to index the collection. We employ machine learning technologies to analyse the visual content of each video segment, and to index it using a vocabulary that contains two major parts: 1,000 semantic concepts (e.g. screen, sky, indoor, pizza... etc.) and around 100 highly discriminative visual objects (brands logos and alpha-numerical characters). This keyword search modality can be used for both Visual KIS and Descriptive KIS search categories.

Semantic Concept Indexing In order to allow users to formulate semantic queries, we employed pre-trained deep convolutional neural networks (CNN) (using the Caffe software [8]) to identify concepts in video sample frames. We use the output judgements of the pre-trained model "CaffeNet" to calculate ranked list results. This model is trained on images from ImageNet [9] (an image database organized according to the nouns of the WordNet hierarchy where each node is depicted by an average of +500 images), and it is able to classify 1,000 concepts.

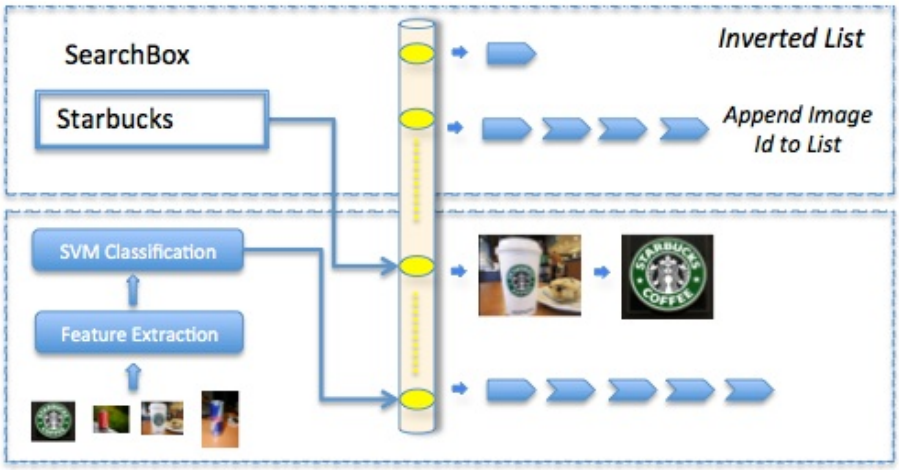

Fig. 1: Inverted Index Structure Based on Semantic Concepts and Visual object Vocabulary.

Visual Object Indexing Unlike the previous approach, which is based on pre-trained models for 1,000 concepts, in this approach we are adding to this list of 1,000 concepts other objects and words extracted from the collection using visual recognition techniques. Here we focus on the occurrence of known visual elements that share the same appearance over different video segments; these elements can be categorised into:

- logos for known pre-selected brands and products; or
- alphanumerical characters.

The identification of these elements is based on extracting HOG features [10] from query objects (that are mined from external visual data sources) the training a linear SVM classifier which is able to identify the occurrence of the query object in the data set sampled key-frames.

For the identification of logos, we are using an approach inspired from the work in [2], Figure 1 shows that the occurrence of *Starbucks Logo* in two images has led into adding them to the same entry (Starbucks entry) of an inverted index for online search, so both of them can be found now by type "Starbucks" as a query (knowing that Starbucks is not among the 1,000 pre-trained concept in 2.1)

The identification of alphanumeric characters is used in order to estimate what are the possible words that exist in a keyframe. For that we are using the approach of [3] that uses an external lexical base and a specific algorithm for candidate words estimation from identified letters and numbers. In Figure 2, we present example of estimated words from keyframes taken from last year query videos. While searching, the user just type an alphanumerical strings that he saw in the query video and the system will search for videos containing these strings.

Fig. 2: An example of identified alphanumeric strings that are used to index keyframes, and thus can be used in the text query to retrieve relevant segments (Images are taken from last year's query videos [1]).Left: *LAAT* and Right: *10, 2010*

All the video segments are indexed using the Semantic Concept and the Visual Object Indexing approaches; both approaches assign to each segment a

score between 0 and 1 indicating the computed degree of presence or absence of a concept or a visual object in the segment. The result is saved in an inverted index file [7] which groups the video segments according to the vocabulary words. During the online retrieval process, a ranking score will be computed based on the decreasing value of presence score.

2.2 Color-based Retrieval

The signature-based video browser tool from Loko et al. [4] performed well in last year's VBS. In their approach, color regions and their position have been extracted from video frames and feature signatures are generated to represent those video frames offline. When interactively searching, users need only to memorise a scene with significant color spot and then draw simple sketches as input to the ranking system. This approach is conceptually simple, yet surprisingly powerful as a method for visual known item search scenario. This approach is successful for two reasons: firstly, it allows users to pick a frame with only few significant color regions by quickly viewing the query video segment; and secondly, it has a interactive interface which help users to produce a simple but flexible sketch query by drawing a few color spots. In this way, we add a color-region based retrieval modality to the system.

2.3 Human Face Filtering

Human face filtering functionality is provided to allow users to filter the videos according to their inclusion of a human faces. This filtering functionality will be be useful for queries involving people and has been considered effective in previous editions of the VBS. To this end, provide a filter check-box to toggle face requirements. This operates as a filter and if checked, the interface will limit the results to videos that contains human faces, all those that do not contain human faces will be filtered out. We use the Viola-Jones face detector [6] to detect human faces in the videos.

3 System Interactive Interface

The user interface is interactive and flexible, which allows a user to produce an effective search query. The interface is designed to employ multiple retrieval modalities. As a standard framework, a text input box allows users to type keyword and a panel displays the sample frames of the top 100 video segments from the ranking system. Users are allowed to scroll left and right to understand the context of each video segment. In addition, the interface includes a canvas panel to help users to quickly draw sketch to toggle the color based retrieval modality. Users are able to choose color, different types of brush and adjust contrast and so on. In addition, the human face filter can be turned on or off for queries which involve people.

4 Conclusion

This paper presents our fourth participation in the Video Browser Showdown. Learning from the past experience and by taking into account both searches categories of this year (Visual KIS and Descriptive KIS), we propose a video browsing and retrieval system that incorporate different retrieval modalities in order to explore the test video collection from multiple angles.

References

1. David Scott, Zhenxing Zhang, Rami Albatal, Kevin McGuinness, Esra Acar, Frank Hopfgartner, Cathal Gurrin, Noel E. OConnor, Alan F. Smeaton. Audio-Visual Classification Video Browser. In MultiMedia Modeling Lecture Notes in Computer Science Volume 8326, 2014, pp 398-401
2. Tomasz Malisiewicz, Abhinav Gupta, Alexei A. Efros. Ensemble of Exemplar-SVMs for Object Detection and Beyond . In ICCV, 2011.
3. Anand Mishra, Karteek Alahari, C. V. Jawahar. Image Retrieval using Textual Cues, In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013.
4. Jakub Loko, Adam Blaek, Tom Skopal. Signature-Based Video Browser, In Proceedings of the 20th MultiMedia Modeling Conference,pp 415-418 , Dublin (January 2014)
5. Lux Mathias, Savvas A. Chatzichristofis. Lire: Lucene Image Retrieval An Extensible Java CBIR Library. In proceedings of the 16th ACM International Conference on Multimedia, pp. 1085-1088, Vancouver, Canada, 2008
6. Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. pages 511518, 2001.
7. C. D. Manning, P. Raghavan, and H. Schutze.introduction to Information Retrieval. Cambridge University Press, 2008.
8. Yangqing Jia. Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding. `http://caffe.berkeleyvision.org/`, 2013.
9. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.
10. Navneet Dalal , Bill Triggs, Histograms of Oriented Gradients for Human Detection, Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1, p.886-893, June 20-26, 2005.