Browsing Linked Video Archives of WWW Video

Zhenxing Zhang[1], Cathal Gurrin[1], Jinlin Guo[1]

[1] CLARITY: Centre for Sensor Web Technologies
Glasnevin, Dublin 9, Ireland
{zzhang, cgurrin}@computing.dcu.ie jinlin.guo2@mail.dcu.ie

Abstract. In this paper, we describe an interactive video browsing system based on a graph of linked video objects. The system automatically organizes unstructured video archives by exploiting visual content similarity between objects in the videos. By generating a video link graph, the system can conceptually groups the videos that contains same objects together for searching and browsing. Both the chosen measures of video object similarity and the video data mining technologies are discussed here and included in the related software demonstrator. In addition, the software offers a query-by-image-example video search capability to jump into the video graph at a certain point to begin browsing the archive.

1 Introduction

Recently, various research efforts have been carried on developing new approaches to object retrieval in large image collections [1]. By applying the efficient text-based query mechanisms, videos containing similar objects can be accurately retrieved from a large dataset in an efficient manner. In this work, we examine new opportunities to study the relationship of videos in a large collection. Instead of using a conventional semantic concept classifier to identify the semantic concepts from videos, our technique links videos based on the presence of objects/feature points within the keyframes of the video content. Automatically linking these videos can support efficient and extensible indexing, fast linkage generation and gives users both links to related content as well as a complete and clear picture of the relationship graph for the entire video collections. Useful information can consequently be summarized together to help users understanding, browsing, and searching of these videos.

2 Indexing

Our aim is to link videos that contain the similar visual entities together, taking into account variances in terms of scale, capture angle, illumination or color appearance. An efficient and accurate indexing and searching algorithm is needed. In this section, the methods and algorithms we employed to address these problems will be discussed in detail.

2.1 Dataset

For this work, a subset of the video collection (about 5,000) for the TRECvid 2012 instance search task was employed, which is composed of user generated videos; hence it is unstructured video data and is likely to have few descriptive annotations. It was originally designed for the task of finding more video segments of a certain specific person, object, or place, given a visual example. This was considered an ideal archive for this prototype video browser for linked archives.

2.2 Video Structure Parsing and Keyframe Selection

In order to automatically organizing a large and unstructured video data collection such as the archive we use here, a shot based segmentation method has been employed to logically divide each video to different shots and one keyframe is selected to represent each shot. This set of keyframes extracted from videos can then be used as still images in large database and will be the subject of the feature extraction and linkage techniques described below.

2.3 Keyframe Representation

By employing a "bag of visual words" model [1], each keyframe is represented as a vector of visual words. For each keyframe in the archive, the affine-invariant Harris-Laplace regions are detected using the technique provided by VLFeat [2]. These regions are the stable areas that are invariant to viewpoint, illumination and scale changes. A 128 dimension SIFT descriptor [3] is then generated based on these regions which will be used in the vector quantization process. We randomly select 20 million descriptors to generate the codebook. The approximate k-means algorithm of [1] has been employed to do the clustering. Each descriptor in keyframes is then assigned with the nearest visual word (cluster center) using the approximate nearest neighbor method. It is important to mention that there will be a quantization error during assignment of descriptors to visual words. Two descriptors that have no similarity could be assigned to same visual words and a link at the keyframe level. This problem will be addressed in detail in next section.

2.4 Link two videos with spatial verification

Based on the visual words, the link between any keyframe and another will be calculated using L2 distance. Using a conventional TF-IDF weighting scheme, we can tune the performance by reducing the contribution of commonly occurring visual words. For each keyframe, we retrieve the top 1,000 ranked keyframes. Random

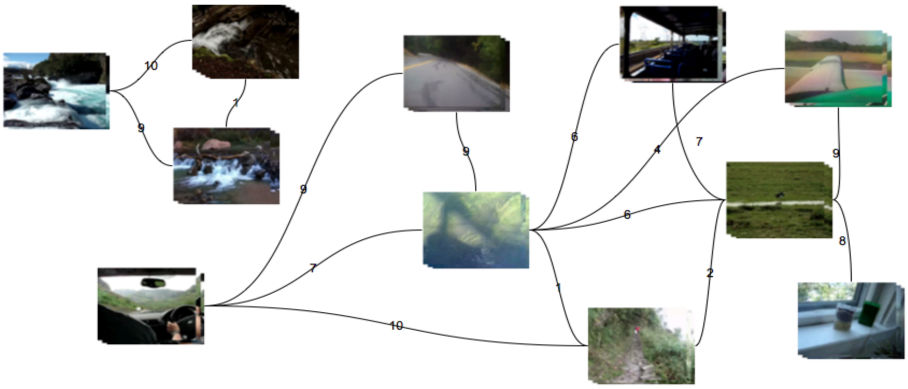

Fig. 1. This figure shows a sample view of the link video graph. Every node in this graph stands for a video from the dataset. The key frames from videos are used for quick view of the video content.

Sample Consensus (RANSAC) [1] algorithm is applied to solve the quantization error by verifying the spatial consistence between frames. False matching visual words will be filtered out because they don't follow the affine transform rule that the true matches will. To implement RANSAC, we randomly choose 4 match pairs to estimate the affine transformation parameters, and do this for 100 times to get the best model. The result of spatial verification can be viewed in the linkage graph presented in Figure 1.

3. The Link Graph and Browser

The video link graph is the index upon which the demonstration software is built. In the video link graph, each node stands for a video in the data collection and weighted edges reflect the (strength of the links) between videos. Multiple links between keyframes in similar videos will result in higher strength links between videos. The browser software is web based to support easy access. To find a good point to begin browsing the archive, the user may submit a photo/image which is processed (as described above for keyframes) to find the most similar videos and this becomes the starting point. When viewing a video, links are provided to the top ranked videos to allow traversal of the linkage graph. Future work involves deploying linkage algorithms (such as a variation of PageRank) over the linkage graph.

References

1. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In Proc. CVPR, 2007.
2. Vedaldi, A. and Fulkerson, B. VLFeat - An open and portable library of computer vision algorithms. ACM International Conference on Multimedia (2010)
3. D. Lowe. Distinctive image features from scale invariant keypoints. International Journal of Computer Vision, 60(2): 91-110, 2004.