DCU at MMM 2013 Video Browser Showdown

David Scott, Jinlin Guo, Cathal Gurrin, Frank Hopfgartner, Kevin McGuinness, Noel E. O'Connor, Alan F. Smeaton, Yang Yang, and Zhenxing Zhang

Dublin City University
Glasnevin, Dublin 9, Ireland
{dscott,jguo,cgurrin,fhopfgartner,asmeaton,yyang,zzhang}@computing.dcu.ie
{mcguinne,oconnorn}@eeng.dcu.ie

Abstract. This paper describes a handheld video browser that incorporates shot boundary detection, key frame extraction, semantic content analysis, key frame browsing, and similarity search.

Keywords: video browser showdown

1 Introduction

In last year's MMM Video Browser Showdown we participated with a handheld based video browsing system that supported two methods of search: concept search and key frame similarity [1]. Building on the experience that we gained while participating in this event, we compete in the next showdown with a more advanced browsing system. This paper provides a short overview of the features and functionality of our new system.

2 Data Processing

2.1 Video segmentation

To ease access to the video data, we segment the video into its shots following Pickering et al. [2]. To avoid including similar frames, we remove duplicates by comparing the global color layout of all key frames within each shot.

2.2 Visual Concept Classification

An important feature of our browsing system is the ability to filter search results based on the appearance of certain concepts such as *persons, vehicles, on screen text,* or *landscapes*. This filtering approach, also referred to as concept-based search, is an effective method to bridge the semantic gap between low-level features and high-level semantics. We trained a set of concept detectors using the popular Bag-of-Visual-Word (BoVW) feature representation and Support Vector Machine (SVM) classifiers.

2.3 Visual Similarity Search

Similarity search is implemented on the key frames by aggregating local descriptors to form a low dimensional global image descriptor. We use sparse SIFT [3] as the local descriptor and perform aggregation using the VLAD method proposed in [4]. This method is a simplification of the Fisher kernel representation [5], but is lower dimensional and less costly to compute. VLAD descriptors will be computed by first performing k-means clustering (with k = 64) on the entire set of SIFT descriptors extracted for each video, then aggregating the differences between each local descriptor in an image and its nearest neighbor, and finally concatenating these local differences to form a $d \times k$ descriptor, where $d = 128$ is the dimension of the local SIFT descriptors. PCA is then used to reduce the dimension of the VLAD descriptors to 128 dimensions. Since this representation allows us to represent all the key frames a video in less than 50MB of memory, we omit the product quantization step and perform search by exhaustively computing the distance between the query VLAD vector for the image and all other key frames in the dataset, rather than using approximate nearest neighbors [6].

2.4 Visual Similarity-Based Clustering

To reduce the cognitive load associated with scanning through a large number of heterogeneous key frames, we implemented functionality that allows the user to group together visually similar key frames and browse the resulting groups. Grouping is achieved by performing agglomerative clustering on the VLAD descriptors extracted for similarity search.

2.5 Face browsing and search

We anticipate that providing functionality to allow users to get a high-level overview of all the human faces appearing in a video will be useful for queries involving people. To this end, we provide a face view that shows all the faces found in the video, and allow users to quickly navigate to the locations in the video in which selected faces appear. We use the Viola-Jones face detector [7] to first locate faces in the videos, and then to cluster these faces by using agglomerative techniques. This clustering also allows face-based search to be easily implemented: when the user chooses to search for similar faces on a given key frame, all images associated with the clusters containing any faces that appear in the key frame are retrieved and displayed.

3 Graphical User Interface

Figure 1 depicts a screenshot of our graphical user interface. It is designed to be used on a tablet PC: either an iPad or any Android tablet. As can be seen in the screenshot, clustered results (shots) are represented by their key frame. On the top of the interface, users can apply different filters and enable similarity search

in the title bar. After tapping (or clicking) on one of the key frames in the result list, a content menu appears and the user has the choice to (a) start playing the video shot, (b) find similar shots in the whole video, or (c) find shots containing similar faces.

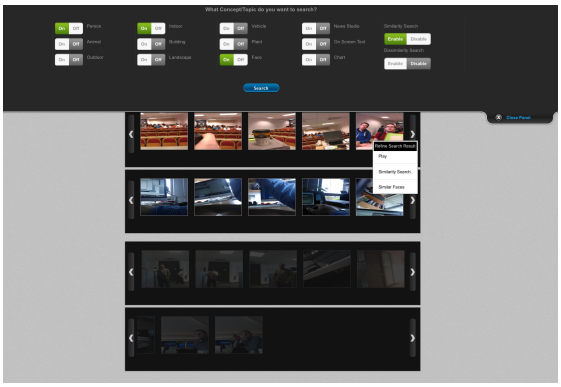

Fig. 1. Screenshot of the Video Browsing Interface.

Acknowledgments

This research was supported by the Norwegian Research Council (CRI number: 174867), Science Foundation Ireland under Grant No.: 07/CE/I1147 and by the EU FP7 Project AXES ICT-269980.

References

1. Scott, D., Guo, J., Wang, H., Yang, Y., Hopfgartner, F., Gurrin, C.: Clipboard: A visual search and browsing engine for tablet and pc. In: MMM. (2012) 646–648
2. Pickering, M.J., Rüger, S.M.: Evaluation of key frame-based retrieval techniques for video. Computer Vision and Image Understanding **92**(2-3) (2003) 217–235
3. Lowe, D.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision. (1999) 1150–1157
4. Jègou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2010) 3304–3311
5. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 1-8. (2007)
6. Jègou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011) 117–128
7. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001). Volume 1. (2001) 511–518