Clipboard: A Visual Search and Browsing Engine for Tablet and PC

David Scott, Jinlin Guo, Hongyi Wang, Yang Yang,
Frank Hopfgartner, and Cathal Gurrin

Dublin City University,
Glasnevin, Dublin 9, Ireland
{dscott,jguo,yang.yang,frank.hopfgartner,cgurrin}@computing.dcu.ie,
hongyi.wang3@mail.dcu.ie

Abstract. In this work, we present a handheld video browser that utilizes two methods of search; Concept Search and Keyframe Similarity. Concept Search allows a user to define a query using selected visual concepts and presents the user with a cluster of video segments based on extracted image features using OpponentSIFT. Keyframe Similarity has a dependance on the previous search for input criteria, allowing a user to select a keyframe for similarity search, returning three types of results; local keyframes from the current scene, global shot similarity based on visual features and text similarity of shots, based on frequently occurring words generated from ASR transcripts.

Key words: Multi-modal Access, tablet pc, visual concept, keyframe similarity

1 Introduction

Having been involved in TRECVid since its beginnings, participating in Ad-hoc search, Instance Search and most recently in the Known Item Search task [2], we have developed many video browser systems for evaluation. In this work, we present a handheld video search and browsing engine that integrates shot boundary detection, optimal keyframe extraction, scene detection, concept-based querying, keyframe browsing and three-way similarity search to allow a user to locate a known video item with minimum input and in as short a time as possible.

In this work our chosen platform is a tablet PC, which can be either an iPad or any Android tablet. The interface is developed in HMTL 5, thereby allowing cross-platform deployment. The user is presented with an interface which has a title bar (top of the screen, permanently visible) containing a set of concepts that help to partition the collection. Users select concepts to build a visual query, this visual query returns a ranked list of shots which are displayed in descending order of relevance. The user can either select a keyframe to determine if it is correct, select other concepts to refine the search or do a similarity search on the selected keyframe to obtain items with similar visual, textual features or keyframes from the same video segment. At the point when the user has found

the required video segment, s/he will tag the video segment and move onto process another information need. There are a number of key search/browsing techniques that our tablet search engine implements:

- **Concept Search**: Models are trained to recognize real-world entities such as *Person, Vehicle, Building* etc, from the source video. These extracted keyframes are compared to these models and given a probability of containing query defined concept features and the engine ranks the results on this probability in descending order.
- **Keyframe Browsing**: The results returned to the user are in the form of a keyframe browser, in ranked order. The user may browse through these results attained through searching as they will contain the entire collection of keyframes.
- **Similarity Search**: Upon selection of a keyframe that seems similar to the user information need, the user is presented with a tabular view of keyframes within the same video scene, a list of keyframes containing comparative low-level features and a list of keyframes which share similar textual features based on ASR keywords.

In the following section we will discuss in more detail the underlying technologies that support the segmentation, searching and browsing functionality.

2 Technical Components

2.1 Constituent Video Segmentation

In order to facilitate easy browsing through the video, we have implemented a shot boundary detection algorithm and a scene segmentation algorithm. Keyframes are selected to represent each video shot by calculating the most average frame by determining the average vector representation of the MPEG-7 descriptors and finding the closest frame to it. Scene segmentation enables the browsing through an entire scene associated with any selected video shot, which a user can do when exploring shot similarity. This is achieved by making the assumption that unrelated scenes have a significantly different representation of the audio signal. Therefore, we determine the Mel-Frequency Cepstrum Coefficients (MFCCs) of the video's audio layer and segment the video by identifying the neighbouring coefficients which have a delta above a threshold.

2.2 Concept Search

Concept-based search is a effective method to bridge the gap between low-level features and high-level semantics. A series of concept detectors were trained by using a SVM framework and the popular Bag-of-Visual-Word (BoVW) model for keyframe-visual-content representation. More specifically, in the BoVW model, we extract OpponentSIFT feature; then k-means clustering is used for construction of a visual vocabulary with 1024 visual words. Finally, for each keyframe, a

1024-dimension histogram is generated by summing all the occurrences of each cluster (visual word), using the nearest neighbour centroid for each extracted feature. In the SVM, the χ^2 kernel is used since it achieves better performance when comparing with other kernels for concept detection. By employing concept search, a ranked list is generated for each user query that ranks the *entire collection of keyframes* for rapid browsing, hence potentially shots will be highlighted by pre-calculated using the OpponentSIFT features visual descriptors mentioned above. The ranked list is likely to be long, so the user is able to navigate through the keyframes by using swipe gestures.

2.3 Similarity Search

Similarity search is a secondary search technique, which is activated once a user selects a keyframe from the ranked list as form of relevance feedback. Similarity search returns a tabbed view of:

- **Local keyframes** within the video segment in temporal order.
- **Globally similar keyframes** based on a fusion of MPEG-7 descriptors; edge, color histogram and scalable color, which produces a ranked list of keyframes in decreasing order of similarity.
- **Textual similarity** based on precomputed similarity of all shots, calculated using conventional text IR techniques operating on the ASR transcripts of the video shots.

3 Conclusion

In conclusion, this paper presents a visual search and browsing engine for handheld devices. This engine has been designed for situations in which a user needs to locate known items in a short time period. Key features of the engine are shot/scene boundary detection, concept-based search, keyframe browsing and three-way similarity search.

Acknowledgements

The research was funded by iAD - information Access Disruptions, a centre for research-based innovation with CRI number: 174867, funded in part by the Norwegian Research Council.

References

1. Koen E. A. van de Sande, Theo Gevers and Cees G. M. Snoek, Evaluating Color Descriptors for Object and Scene Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 32 (9), pages 1582-1596, 2010.
2. Colum Foley and Jinlin Guo and David Scott and Paul Ferguson and Cathal Gurrin and Alan F. Smeaton. TRECVid 2010 Experiments at Dublin City University, TRECVid 2010 - Text REtrieval Conference TRECVid Workshop, 2010, Gaithersburg, MD.