Eolas: Video Retrieval Application for Helping Tourists

Zhenxing Zhang, Yang Yang, Ran Cui, and Cathal Gurrin

School of Computing, Dublin City University
Glassnevin, Dublin 9, Dublin, Ireland
{zzhang,cgurrin,yang.yang}@computing.dcu.ie
{cuiran1991}@sina.com

Abstract. In this paper, a video retrieval application for the Android mobile platform is described. The application utilises computer vision technologies that, given a photo of a landmark of interest, will automatically locate online videos about that landmark. Content-based video retrieval technologies are adopted to find the most relevant videos based on visual similarity of video content. The system has been evaluated using a custom test collection with human annotated ground truth. We show that our system is effective, both in terms of speed and accuracy. This application is proposed for demonstration at MMM2014 and we are sure that this application would benefit tourists either planning travel or while travelling in real-time.

Keywords: Multimedia Information Retrieval, Video Processing, Exemplar-SVMs, Visual Similarity

1 Introduction

The motivation of this work is to help tourists automatically finding documentary videos concerning a landmark of interest. The proposed query mechanism is via a photograph, either captured in the moment, or chosen from the photo-album. This application would be extremely helpful especially when they travelled in foreign countries with different languages. Recent advances in content-based video retrieval research suggests that is now possible to apply robust and efficient techniques to solve some real world problems; in this case, the problem of 'finding out about' certain landmarks. There has been some prior work in this area, such as using a photo to identify certain classes of objects [4], location recognition from captured images with a mobile device [5], or automatically identifying a sculpture [1] and labelling it and so on. In this demo paper, our purpose is not only to develop a novel application for tourists, but also to bring state-of-the-art video retrieval technologies beyond desktop environment into a real-time mobile devices usage.

In this demonstration we present video retrieval system, named Eolas (the Irish word for 'eyes'), that we implemented using exemplar-SVMs based object detection technologies [3]. The focus of this implementation is on linking famous landmarks/attractions with documentary videos from online sources such as YouTube. By compare the visual similarity between a selected query image and keyframes extracted from a video archive, the system is able to present a list of documentary videos which are most closely visually-related to the query image.

In the rest of this demo paper we first provide an overview of retrieval system, then present the implementation details and system performance, and finally we present some concluding remarks and suggestions.

2 System Overview

Eolas is composed of two main components, a smartphone application and a online web service. Even though smartphones have evaluated dramatically with both computational and storage capabilities, they are still not ideal for heavy processing tasks such as video processing and retrieval. Hence, we use the online web service based on a remote server to accomplish efficient content analysis and retrieval, with the smartphone simply being the user interaction tool.

Online Search Architecture The online query processing engine is triggered when the smartphone transmits a photo (from the Eolas applicaiton via the camera or the photo-album). Prior to transmission, there is phase of initial filtering and encoding which optimises the photo for upload. The video retrieval service that received the query has three main components: *a*) Query Parsing Module which responsible for parsing a query request, extracting feature representation of query image and passing the result to retrieval engine. *b*) Retrieval Module which performs searching operation based on the pre-indexed dataset, and return the ranked results to client side in JSON format. *c*) Server Log Module which can save the service operations related to each query.

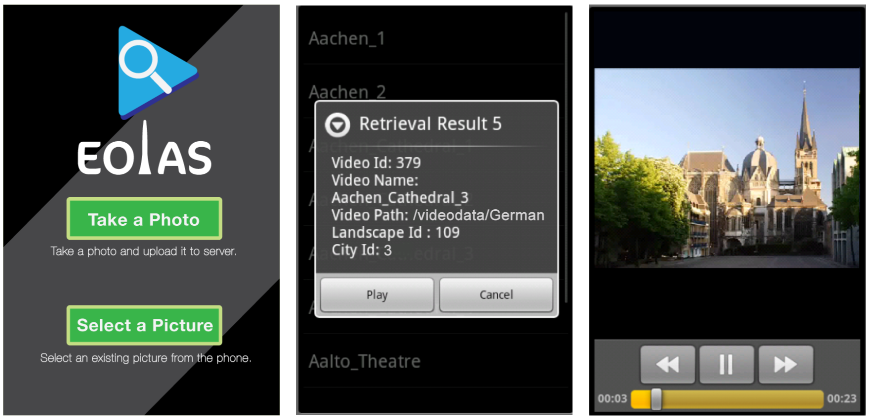
Mobile Application GUI The screenshots of smartphone application are displayed in Figure 1. Users can take a photo to query. A ranked results will be displayed and more details can be presented after click any result.

3 System Implementation and Performance Evaluation

Based on the previously described architecture, we choose the tourist application as an initial use-case. We constructed a dataset of famous German tourist attractions. 310 documentary videos were downloaded from Flickr website under the Creative Commons license. The videos average at two minutes in duration and each one focuses on one attraction. The server application indexed these videos as follows.

3.1 Offline Video Processing Pipeline

There are three main steps in the offline video processing pipeline. Firstly, in order to reduce the complexity, each video has been segmented into a series



Video Id: 379
Video Name:
Aachen_Cathedral_3
Video Path: /videodata/German
Landscape Id : 109
City Id: 3

(a) (b) (c)

Fig. 1: snapshot of smartphone application GUI

of shots using a conventional approach to shot boundary detection (background colour change) and each shot is represented by one (central) keyframe. This gives 2,610 keyframes. Secondly, Histogram of Oriented Gradients (HOG) are picked to describe each keyframe, due to its good performance for object detection, robustness, and its speed. More importantly, it could offer stable performance even when there are many local changes, such as illumination changes, where local descriptors like SIFT would normally fail to match. Finally, a relational database has been used to index the meta data and the feature representation.

3.2 Exemplar-SVM based Scoring Scheme

Different from approach of [7] using a bag of visual word representation and a text-based indexing algorithm, we implemented a linear discriminative object classifier for each query image, as was done by [3], [6]. There are two major benefits of this approach, firstly unlike [7], there are no quantization error because we are using the Hog descriptors directly, and secondly, a unique weighting score can be learned by using this data-driven learning method. This allows us to determine the most discriminative visual features according to one positive query example and many negative examples. After obtaining a weighting vector $\overrightarrow{w}$ for a query image, each video can be sorted using the following technique:

$$S(I_q, I_i) = \overrightarrow{w}^T X_i \tag{1}$$

where I_q is the query image, I_i is the ith video and X_i is the feature vector. The top ranked videos are returned to the user. About 10,000 random images have been downloaded from Flickr website matching common topics like human, tree, parties. These images are checked and then used as negative examples for every training process which can provide for fast online retrieval. *LIBLINEAR* library [2] has been employed to achieve fast online training.

3.3 Performance Evaluation

Ten topics have been manually annotated as ground truth and in a user study, we calculated the mean average precision (mAP) and average query time across all topics (shown in Table 1). Eolas is shown to return 69% true positive results in less then 15 millisecond.

Table 1: System Evaluation Performance

Dataset Size		Performance	
Videos	Keyframes	mAP	Query time
311	2610	0.69	0.014 (s)

4 Conclusion

This demonstration paper presents a real-world application of a content based video retrieval engine using machine learning technologies. We describe the system and provide evaluation results in a small user study. Further work would be to expand the data to include different locations, to include GPS data from the image EXIF header to filter potential videos to a region, and to extend this work to different use-cases.

References

1. R. Arandjelović and A. Zisserman. Name that sculpture. In *ACM International Conference on Multimedia Retrieval*, 2012.
2. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
3. Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.
4. Google Mobile. Open your eyes: Google goggles now available on iphone in google mobile app. `http://googlemobile.blogspot.ie/2010/10/open-your-eyes-google-goggles-now.html/`, 2010.
5. Georg Schroth, Robert Huitl, David Chen, Mohammad Abu-Alqumsan, Anas Al-Nuaimi, and Eckehard Steinbach. Mobile visual location recognition. *IEEE Signal Processing Magazine, Special Issue on Mobile Media Search*, Vol. 28, No. 4, pp. 77-89, 2011.
6. Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Transaction of Graphics (TOG) (Proceedings of ACM SIGGRAPH ASIA)*, 30(6), 2011.
7. J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003.