Real-time Query-by-Image Video Search System

Andre Araujo, David Chen, Peter Vajda and Bernd Girod
Department of Electrical Engineering, Stanford University, Stanford, CA 94305, U.S.A.
{afaraujo, dmchen, pvajda, bgirod}@stanford.edu

ABSTRACT

We demonstrate a novel multimedia system that continuously indexes videos and enables real-time search using images, with a broad range of potential applications. Television shows are recorded and indexed continuously, and iconic images from recent events are discovered automatically. Users can query an uploaded image or an image in the web. When a result is served, the user can play the video clip from the beginning or from the point in time where the retrieved image was found.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

query-by-image, trending images, video indexing, video search

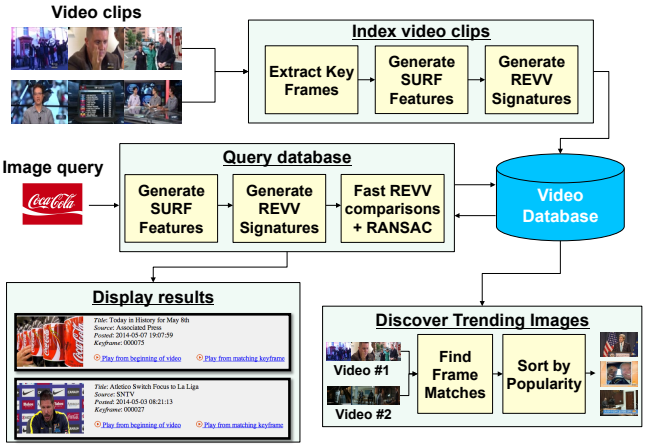

Figure 1: Overview of our query-by-image video search system.

1. INTRODUCTION

Despite the widespread adoption of visual search systems in recent years, a large body of visual content, in the form of videos, cannot be searched using a query image with today's commercial systems. There has been research in this area [1, 6, 7], but no practical system has emerged. We introduce a system that continuously indexes new videos and allows search using images.

This type of technology has a diverse set of potential applications. For advertisement monitoring, users can find all points in time in a video where a particular logo or object was shown. In education, users might use a slide to find the point in time in a lecture video where a certain concept was explained [4]. For content linking, users could use an image in an article to find related video stories. Or they could snap a picture from a display showing a video to get more information about it.

There are, however, several challenges. Videos present an enormous amount of data: the memory usage and search latency for such a system can be significant. To address both of these issues, we use a scheme based on the Residual Enhanced Visual Vector (REVV) [3], a state-of-the-art image retrieval technique. In this technical demo, we present our live, real-time query-by-image video search system that can index a large database of news videos. Please see our demonstration video[1] for additional information and also visit the system's website[2].

MM'14, November 3–7, 2014, Orlando, Florida, USA.
ACM 978-1-4503-3063-3/14/11.
http://dx.doi.org/10.1145/2647868.2654867.

2. SYSTEM ARCHITECTURE

Figure 1 presents a block diagram of our system. The main blocks are described in the following subsections. We use the system front-end developed for the *Eigennews* project [5]. A diverse set of television news programs from 59 channels are recorded, and the videos are analyzed using a multimodal approach [5]. Around 500 video clips are generated per day, and their average duration is 3 minutes.

2.1 Index video clips

For each video clip, we extract keyframes at 1 fps. SURF features [2] are extracted from each keyframe and aggregated into a REVV global signature [3]. The REVV signature enables very memory-efficient indexing and fast retrieval. Each keyframe's REVV signature is stored in memory for fast querying, while the keyframe's SURF descriptors are stored in disk – they will be used in the geometric verification step, which is performed for a small number of keyframes.

Video indexing happens continuously in the background, updating the database with the most current news programs. We use a predefined threshold on the number of keyframes the system can index, in order to limit memory utilization. When a new video is to be included in the database, we check if this threshold is surpassed. If it is, the oldest video in the database is removed before the new video is included. Our memory limitation corresponds to 1 million keyframes, which allows for about 10 days of news videos to be indexed

[1] http://blackhole1.stanford.edu/vidsearch/VideoSearchDemo.mov
[2] http://videosearch.stanford.edu

Figure 2: Examples of trending images detected on May 6, 2014.

in the system. This way, our system always indexes news programs from the most recent 10 days.

After a 30 minute news program is uploaded, the system takes on average 15 minutes to segment it into individual stories (generating one video clip per story), relate the stories to each other and extract metadata [5]. After that, the process from keyframe extraction to including the new REVV signatures in the database takes on average 5 minutes. In our system, each REVV signature utilizes 190 centroids and our system needs to store on average a 512-byte REVV signature per keyframe. With the 1 million keyframe budget, our system needs about 500 MB to index news programs from the past 10 days.

2.2 Query database

Once an image query is submitted, the system immediately extracts SURF features and then generates a query REVV signature. The query's REVV signature can be compared to the database's REVV signatures using bitwise operations [3], which speeds up this step. This process generates a ranked list with the most similar keyframes (according to the REVV comparisons) on top.

After this initial pass over the database, we go over the generated list and match some of the top keyframes to the query image, based on their SURF features. In this process, we first find pairs of query and database features that are sufficiently close, and then try to find an affine transformation between them using RANSAC. We consider a keyframe to be a match if there are at least 8 inliers.

In this process, keyframes that are close in time tend to be ranked closely in the list generated using REVV signatures. During the geometric verification step, we avoid checking keyframes that are within 5 seconds of previously checked keyframes, to avoid redundant computation. The geometric verification step checks 100 keyframes and this near-duplicate removal strategy allows for diverse videos to be compared.

Other system tasks during query time include retrieving the video information in a database, fetching and converting the image if necessary. On average, results are served 7 seconds after a query is initiated, using one core on an Intel Xeon 2.4GHz processor.

2.3 Discover trending images

Once a new video is indexed in the database, we match it against videos that were uploaded in the most recent 24 hours to find trending images. These are iconic images that were broadcast by multiple channels, and usually correspond to important events. Each new video is matched against recent videos only from different channels, in order to find images that were broadcast by different news sources. The matching process uses a similar approach to the one described in the previous subsection, using a combination of REVV and SURF matching, and takes on average 81 seconds per video.

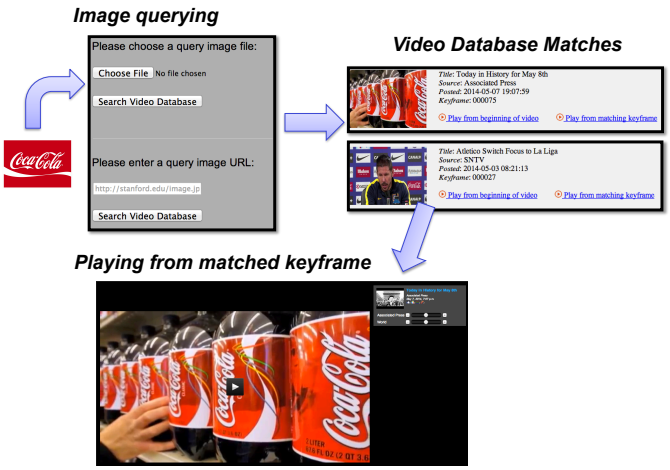

Figure 3: An example of the user interaction with the system.

Figure 2 shows detected trending images from May 6, 2014. They find stories related to the kidnapped Nigerian girls, the World Cup, Ukraine crisis, a tragic accident, climate change and Hillary Clinton's presidential ambitions, all among the top news stories covered in that day.

3. USER INTERFACE

Figure 3 presents a sample of the webpages that the user can access during an interaction with the system. From the system's homepage, an image can be uploaded from the user's computer, or by pasting the URL of an image from the web. Suggested queries and trending images are also presented in the system's webpage. Once the user submits an image query or clicks on a suggested image, the results page is presented. The matched keyframe is displayed for each match, along with detailed information about the program/channel name, clip title, date/time and exact keyframe in the video clip. The interface further allows the user to play the discovered video clip within the *Eigennews* system, either from the beginning or from the exact point in time where the matching keyframe occurred.

4. REFERENCES

[1] A. Araujo, M. Makar, V. Chandrasekhar, D. Chen, S. Tsai, H. Chen, R. Angst, and B. Girod. Efficient video search using image queries. In *Proc. ICIP*, 2014.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. *CVIU*, 110(3), 2008.

[3] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod. Residual Enhanced Visual Vector as a Compact Signature for Mobile Visual Search. *Signal Processing*, 93(8), 2013.

[4] N.-M. Cheung, D. Chen, V. Chandrasekhar, S. Tsai, G. Takacs, S. Halawa, and B. Girod. Restoration of Out-of-focus Lecture Video by Automatic Slide Matching. In *Proc. ACM Multimedia*, 2010.

[5] M. Daneshi, P. Vajda, D. Chen, S. Tsai, M. Yu, A. Araujo, H. Chen, and B. Girod. EigenNews: Generating and Delivering Personalized News Videos. In *Proc. IEEE BRUREC*, 2013.

[6] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot. TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proc. TRECVID*, 2013.

[7] J. Sivic and A. Zisserman. Video Google: Efficient visual search of videos. *Toward Category-Level Object Recognition*, 4170, 2006.