Sensor-rich Video Exploration on a Map Interface

Beomjoo Seo, Jia Hao, Guanfeng Wang
School of Computing
National University of Singapore
Singapore 117417
[seobj,haojia,wanggf]@comp.nus.edu.sg

ABSTRACT
Result presentations from searches into video repositories is still a challenging problem. Current systems usually display a ranked list that shows the first frame of each video. Users then explore the videos one-by-one. In our recent work we have investigated the fusion of captured video with a continuous stream of sensor meta-data. These so-called sensor-rich videos can conveniently be captured with today's smartphones. Importantly, the recorded sensor-data streams enable processing and result presentation in novel and useful ways.

In this demonstration we present a system that provides an integrated solution to present videos based on keyframe extraction and interactive, map-based browsing. As a key feature, the system automatically computes popular places based on the collective information from all the available videos. For each video it then extracts keyframes and renders them at their proper location on the map synchronously with the video playback. All the processing is performed in real-time, which allows for an interactive exploration of all the videos in a geographic area.

Categories and Subject Descriptors
H.3.5 [Information Storage and Retrieval]: On-line Information Services—Web-based services; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Sensor Fusion

General Terms
Algorithms, Measurement, Performance

Keywords
Web services, hotspot estimation, keyframe extraction

1. INTRODUCTION
Due to the technological advancements and the popularity of smartphones it is now easy for users to capture, upload and share videos. A number of video hosting services cater to this popular trend. In fact, the explosive growth in mobile video services has already led to their significant contribution to the overall Internet traffic [2].

Some of these videos are associated with geographic locations (i.e., GPS coordinates) and existing hosting solutions make them available through map services for browsing and viewing. However, conventional approaches are limited in that videos are associated with just a single GPS location. Hence, in terms of presentation there exists only a minimal connection between the map interface and the videos that typically have been retrieved based on a user’s query. To explore techniques that provide a more integrated and interactive experience, we have developed a video hosting system called GeoVid (http://geovid.org) [1]. In our novel and unconventional framework, videos and their corresponding meta-data streams are efficiently indexed and synchronously presented on a common map interface.

As a key feature of our demonstration we make use of our recently designed distance-aware video summarization algorithm. Different from existing keyframe extraction and summarization techniques [3, 4], our method processes and utilizes the video meta-data (specifically compass and GPS information) which makes it efficient and light-weight. In our methodology, we are conscious of the fact that the location of the most salient objects in a video are often times not at the position of the camera, but may in reality be quite a distance away (e.g., a video of the Niagara Falls). Clearly it would be very beneficial to place extracted keyframes in their correct geographic locations on a map. In the past, an automatic detection of popular points-of-interest was difficult to achieve1. Our approach provides a keyframe-distance compution and placement with no manual intervention and as such we believe it provides high utility for end-users.

2. SYSTEM ARCHITECTURE
Our acquisition software is implemented as smartphone applications available for both Android and iOS. The application software captures 720 × 480 resolution MPEG-4 videos and also records sensor measurements every 250 milliseconds. After a recording the user can upload the video through a network connection (either WiFi or 3G) to our demonstration server.

The hotspot estimation and keyframe extraction is performed on our demonstration server. After receiving a video and the corresponding sensor measurements (in JSON format) the server processes the GPS location and the view angle of every video frame and inserts them into a spatial database to make the information searchable. To enable efficient wide-area access we transcode the videos to a 768 kbps bitrate and 480 × 360 pixel resolution. A specific back-end process – depicted in a gray box in Fig. 1a – constructs the

1We do not use the term landmark since our technique is useful for a broad set of geographically related phenomena.
FOV (field of view) of every frame and updates the region that is affected on a map grid incrementally. Local maxima throughout the grid are selected as hotspots. Based on the set of hotspot locations the process computes the distances to the camera locations and determines the best-matching hotspot within each FOV as its visible distance. Subsequently, the estimated visible distance provides the input for the keyframe decision. Finally, the computed keyframes are actually retrieved via the image extraction module.

3. DEMONSTRATION

Our demonstration system allows users to experience sensor-rich videos based on the underlying key mechanisms described in the previous section. Fig. 1b illustrates the system interface. A list of video thumbnails are presented along the top and a Google Maps interface is shown in the lower part. The thumbnail list is automatically updated whenever a user navigates across the map. When the mouse pointer hovers over one of the thumbnails, the corresponding GPS trajectory and keyframe related information is asynchronously retrieved from the server and displayed on the map interface. When the user clicks a thumbnail, a map-overlaid player is launched and the video played at the designated starting location. During the playback of the video, the camera’s current location and FOV are animated along the corresponding GPS trajectory, using the Raphael\(^2\) vector graphics engine. If a keyframe is encountered during playback, its thumbnail is positioned at its estimated location on the map along with a fade-out visual effect.

Our prototype system is also useful in identifying some interesting issues. For the keyframe extraction algorithm we observed that a few keyframes do not match well. We found that some popular places tend to have several distinct attractions which are challenging to distinguish. One possible enhancement would be to also leverage existing landmark information. An additional challenge is to provide correct distance estimations in the presence of obstructions such as large buildings. This issue can be reduced if 3D building models are available to check for occlusions.

4. CONCLUSIONS

We presented a novel and integrated video exploration approach where keyframes are positioned at an expected target location during playback on a map interface. Keyframes and their locations are computed in a fully automated manner. Thus, a number of visual cues are provided to the user to effectively navigate a large set of videos.

From our prototype system we have identified a number of areas for improvements and future work. First, there currently exists a slight asynchrony between the video and FOV renderings (typically less than one-second). This is caused by the difficulty of determining the exact time when the first video frame is recorded on the mobile device. Second, the GPS and compass signals are quite noisy and jittery, since we render the raw sensor data without any post-processing. To solve these issues, we plan to apply appropriate low-pass filtering mechanism combined with motion-vector analysis.

We plan to open our APIs for access to all sensor measurements and for users to submit their own keyframe extraction algorithms and hence visually evaluate their quality through our web browser.

5. REFERENCES


\(^2\)http://raphaeljs.com