Searching for Relevant Video Shots in BBC Rushes Using Semantic Web Techniques

Bradley P. Allen[1)] and Valery A. Petrushin[2)]
[1)] Siderean Software, Inc.
[2)] Accenture Technology Labs

Abstract

In this paper we describe an approach that merges Semantic Web technologies (RDF, SKOS) and content based multimedia information annotation and retrieval techniques (MPEG-7). We demonstrate our approach with a system that allows a user to discover and select pieces of relevant information from video clips. The system uses both textual metadata such as duration of video, producer's name, main title and subtitle, copyright owner, keywords, production and broadcast dates, etc., and MPEG-7 low-level visual color and texture features that are extracted from key frames of each shot. The visual features are clustered using self-organizing maps and the centroid key frames are selected as visual "words". The relationships among visual words are represented using the RDF and SKOS standards and are used, together with textual metadata, for efficient faceted navigation.

1. Introduction

In broadcasting and filmmaking industries "rushes" is a term for raw footage, which is used for productions such as TV programs and movies. Not all raw footage goes into a production. A typical "shoot-to-show" ratio for a TV program is in the range from 20 to 40. It means that up to 40 hours of raw footage is converted into one hour of TV program. Watching other people's rushes is boring. They have a lot of static scenes, redundant episodes and out of focus fragments. That is why using rushes for creating TV programs manually is hard and inefficient. However managers believe that rushes could be valuable if technology could help program makers to extract some "generic" episodes with high potential for re-use. Such generic footage is called "stockshots". There are commercial stockshot libraries which provide shots for particular geographical locations, vehicles, people, events, etc. So, the technical problem is how to mine rushes for golden nuggets of stockshots.

Currently, there are several digital asset management systems (for example, *Arkemedia* by Harris Corp. [1]). These systems allow manually providing some data about the video clip, automatically split the clip into shots and select key frames for each shot. These systems are very helpful for archiving media data, but are not powerful for searching for useful shots. On the other hand, there are many experimental systems for news search, annotation and summarization (e.g., Informedia [2]). These systems are heavily based on textual information that comes from close captions or speech transcripts. In rushes textual information is rather sparse – several keywords related to the whole clip and some metadata about the creators of the clip, geographical location, and date of production. Rushes' soundtracks can be noisy and indecipherable for automatic speech recognition. Moreover, the soundtracks of stockshots are rarely used. They separated from the video stream and substitute by another soundtrack. These peculiarities of rushes and stockshots require developing different data mining techniques that combine sparse textual data with dominant visual data.

In this paper we present a Web-based system that helps TV program makers select relevant shots from a repository of shots that were created automatically from rushes. The program maker can use both textual metadata and visual "words" for search.

2. Data

The data set is the BBC Rushes, which consists of 615 video clips that present raw footage that could be reused for creating TV programs. Each video has metadata that includes such items as duration of video, producer's name, main title and subtitle, copyright owner, production and broadcast dates; tape ID where the video is stored, topic number, etc. The metadata also includes the list of keywords and the description of the video clip structure. The video clip is split into "chunks" or shots. Each shot has starting and ending time stamps, and has one or more key frame images that are presented as JPEG files in a separate directory.

If only one key frame represents a shot, then it is usually a frame taken from the beginning or the end of the shot. The number of shots per a clip ranges from 2 to 496. The total number of shots is 10064. The number of key frames per shot is in the range from 1 to 377. The number of key frames for each video clip varies from 2 to 1333. The total number of key frame images is 39,142. The size of each image is 176 by 144 pixels. It seems that metadata on shots were obtained using a semi-automated tool. The quality of shot and key frame extraction is rather poor – some shots are too long and have many genuine shots inside, key frames are often selected from the very beginning and the end of a shot and do not reflect the real content of the shot. However, in spite of it being tempting to redo shot and key frame extraction, we decided to use them "as is" for two reasons: first, to have an opportunity to compare our results with results of other researchers, who are using the original shot segmentation, and, second, to see how our approach works for real industrial data.

3. Proposed solution

The general idea is to use Semantic Web techniques to represent relationships among both textual and visual metadata of various types. Each type of data forms a facet with its own ontology. The relationships among resources (concepts, objects) are described using the Resource Description Framework (RDF) [3] or some tools, such as the Dublin Core (DC) and the Simple Knowledge Organization System (SKOS) [4] that are based on RDF and RDFS representations.

3.1. Metadata representation

The following textual metadata were selected from the metadata that are provided with video clips: main title, subtitle, producer, production date. Clip duration, tape ID, topic number and copyright owner have also been considered as informative and potentially useful asset metadata. These are encoded as Dublin Core attributes occurring in descriptions of individual clips. Shots are related to the clips from which they have been extracted using the *dcterms:partOf* attribute.

Subject metadata is available on a per-clip basis in the form of a description (a set of keywords.) Subject metadata is represented as a set of SKOS concepts, related to shots using the *dc:subject* tag. Concepts that are synonymous with terms that are values of concepts in the Library of Congress Thesaurus of Graphical Materials 1 [5] are represented using TGM-1 concepts. This allows them to be viewed and selected in a faceted navigation interface using the hierarchy defined by the broader term/narrower term relationships between thesaurus concepts.

For visual metadata the low level facets are color, texture and shape. Currently, we take into account color and texture, and leaving shape for future extensions. A number of visual features can be used for describing color and texture of the key frames. We used the following MPEG-7 descriptors [6]: for color – dominant color, color structure, and color layout; for texture – homogenous texture, and edge histogram. The above mentioned features have been extracted for each key frame using the MPEG-7 XM tools. Then Self-Organizing Map (SOM) clustering has been applied for each feature. After human evaluation of clustering results the following features were selected: color structure for representing similarity by color and homogenous texture for representing similarity by texture. Three SOM clusterings have been produced using the above mentioned selected features and their combination. For each map key frames that are closest to the SOM nodes' centroids form "visual words". Thus, we obtained three sets of "visual words" that capture the relationships among key frames by color, texture, and color + texture. Each set contains about 1,000 items. Each node of the SOM is represented as a SKOS concept, with the value being the image associated with the node. Each shot is related to the concepts associated with the nodes that its key frames are members of using the *dc:subject* attribute.

3.2. User interface

A user interface for being useful for a program maker should have means for:

- Navigation over the shot database using a combination of facets derived from textual and visual metadata.
- Selection and manipulation of relevant shots found during the user session.
- Saving the results of a session in a form that can be useful for further processing or usage.

The BBC Rushes search user interface is built as an HTML page using AJAX [7], communicating via HTTP calls with an instance of the Seamark Navigator system [8]. Seamark Navigator is used to store the textual and visual metadata describing clips and extracted shots, and provides SOAP services that support faceted navigation over the metadata, with facets defined using textual and visual attributes of the clip and shot metadata. Seamark provides both free text querying and facet value selection for user search of the shot metadata repository.

The basic unit for navigation is a shot. Each clip in the BBC Rushes may have from 2 to several hundred shots. The textual metadata that is assigned to a clip extends automatically to all its shots. The further refinements can be done based on visual similarity of shots.

Navigation is provided in two manners: through the ability to enter free text search queries against the textual facets, and through the ability to left click on textual and visual facet values. In either manner, the system updates the user interface with a set of search results and a new set on faceted navigation options. The initial state of the interface shows an overview in terms of facet values of the most frequently occurring facet values per facet across the entire collections of shots.

The results of a search are represented as a sequence of shots sorted in accordance with criteria combination of free text queries and facet value selections. Information related to every shot includes data related to the clip (title, subject, producer, etc.) and data related to the shot (start/end time, duration).
Left clicking on any key frame or on a title hyperlink launches a video player for playing back the shot. Shots in the results that are of interest to the user can be copied to a storyboard using drag-and-drop.

The storyboard contains the total number of shots, total duration of shots, and a list of selected shots with key frames and attributes (title/produces/duration, etc). Left clicking on the key frame image runs a video player for the shot. The user can reorder and delete shots on the storyboard using drag-and-drop.

Pressing on “Play All” button invokes an HTTP call to a REST Web service that takes the URLs of selected shots and generates a SMIL [9] document composing the shots into a single virtual clip that allows the user to watch the shots played sequentially in the order as they are listed., and then launches a player that plays the virtual clip.

3.3. User Interface extensions

Taking into account that the collection of video clips is devoted to tourism and vacations and presents clips from various parts of the world some attractive extensions of the user interface could be suggested.

1. Using a map of the world that allows the user seeing how the results of the search are geographically distributed.
2. Using a hierarchy of clusters or self-organizing maps created based on visual features as a hierarchical visual facet for browsing clips and shots.
3. Allowing the user to manage and create clip metadata in the context of faceted navigation.

References

[1] Arkemedia digital asset management system: http://www.broadcast.harris.com
[2] Informedia: http://www.informedia.cs.cmu.edu/
[3] RDF: http://www.w3.org/RDF/
[4] SKOS: http://www.w3.org/2004/02/skos/
[5] TGM: http://www.loc.gov/rr/print/tgm1/toc.html
[6] B.S. Manjunath, Ph. Salembier, Th. Sikora (Eds.) Introduction to MPEG-7. Multimedia Content Description Interface. John Wiley & Sons, Ltd., 2002, 371 p.
[7] AJAX: http://en.wikipedia.org/wiki/AJAX
[8] Seamark: http://www.siderean.com/Seamark_datasheet.pdf
[9] SMIL: http://www.w3.org/TR/REC-smil/