VideoGIS: Segmenting and indexing video based on geographic information

Toni Navarrete, Josep Blat
Departament de Tecnologia, Universitat Pompeu Fabra
Passeig de Circumval·lació, 8 - 08003 - Barcelona (Spain)
Tel: (+34) 93 542 22 37
Fax: (+34) 93 542 24 51
{toni.navarrete, josep.blat}@tecn.upf.es

Abstract The VideoGIS project aims at combining video and geographic information in order to dynamically generate hypervideos that will be navigable by geographic content. After a brief survey, we describe our segmentation approach, based on geographic content, and discuss indexing of both geographic and video information, where we have used Frame-segment trees. Then we describe our geographic related queries and the composition of the resulting video and we discuss the lack of navigability of the current approach and the need of better semantic models of video in order to improve it.

Introduction

Our goal is the construction of dynamic hypervideos based on geographic content with a strong navigational aspect. The aim is, roughly, to define a system connecting geo-referenced video material (a typical case is video shot from the air, where the spatial coordinates of the sequence are known) with the geographical information of this space. Elements of interest (for instance schools, hotels, rivers or county boundaries) can be overprinted on the images. The result is a simple map where the bottom layer is a moving real image. This is what we have called videomap. The system will also be able to generate itineraries depending on the geographic information (for instance “itinerary over the fields of almond trees in a county” or “itinerary over the villages whose income is lower than some threshold”). This is what we have called videoitinerary. This means that the result is a hypervideo, created on-the-fly, and navigable by geographical content. We set it in the framework of what we have called the VideoGIS project.

Our hypothesis is that the use of videomaps and videoitineraries will provide an alternative interface to non-skilled users to traditional maps, which present a higher level of abstraction. This fact would make them suitable for public participation decision processes in the field of urban planning or environmental education.

Our current system prototype splits the video into geographically meaningful units, using a novel segmentation approach, based on coordinates and geographic features. The segments are indexed by means of a frame-segment tree in order to allow queries related to the features found in segments. The result is a dynamic video which can be watched within a web browser. Java [Java URL], XML [XML URL], PostgreSQL [PostgreSQL URL] and Quicktime [QuickTime URL] are used.

The prototype shows that the results can be difficult to manage by the users if dealing with a large video base. We discuss two basic issues to improve it, namely in the semantic modelling and through an improved navigation structure.

In the paper we present first a brief survey analyzing projects combining video and geographical information (GI) and how our approach differs from them. Then we describe the system, and finally we present some conclusions and perspectives for future work.

A brief survey on video and GI

The combination of geographic information and hypermedia has not been widespread until recent years. But even now, videos are small clips used to give a better idea of a concept, to support an explanation, but they are not in the core of the hypermedia system. They are usually atomic, sequential, static and far from being navigable. The only exception is the Aspen Project, [Lippman 1980], [Negroponte 1995], developed at MIT in 1978, curiously the first project on hypermedia. Using four cameras mounted on a truck, the streets of Aspen were filmed (in both directions), taking an image every three meters. The system consists of two videodisk players that allow users to “drive” through the city. However it is made from a closed and well-known set of videos, while our approach deals with an open base.

Other important project from a historical point of view is the BBC Domesday Project [Openshaw et al. 1986], which presents a map of Great Britain that allows users to watch video clips, aerial photographs and listen to
natural sounds from certain places. In this case, videos were only static clips. A lot of other projects have used video clips in a similar way. The most typical case is multimedia atlases where the user can find video clips showing locations or providing a deeper definition of geographical concepts. Probably the best known (at least most sold) is Microsoft Atlas Encarta. Some of these atlases can be found on the web. The *Ailes de les Illes Balears* [UIB 1999], [Blat et al. 2001] was probably the first multimedia atlas on-line.

Other applications with a geographical background have used video clips. [Shiffer 1992] presents a collaborative hypermedia tool for urban planning. In this scenario, a video clip showing the traffic on a highway at different times, or the reproduction of the noise produced by traffic, can help planners to take decisions. An on-line version was also developed [Shiffer 1995]. Nowadays it could take advantage of web cams to show real-time videos. A related project is the CD-ROM of ParcBIT [Blat et al. 1995], a hypermedia application to help architects developing an environmental aware planning for a technology park.

The environmental application [Ferreira 1999] uses video with quite a different approach. The aim is to analyse the emission of pollutants from the stacks of a power plant in Setúbal (Portugal). The video image is processed to calculate the direction of the plume, i.e. the direction of the wind, which is the base for the environmental model. Remote sensing is a field where the use of video has been very intensive, because, as Ferreira explains, it has two advantages over traditional techniques: images can be obtained in real time and video cameras have a high sensitivity. But again, there are no navigational concepts in this type of projects.

[Christel et al. 1999] describe a project in the framework of the Informedia digital library, where geographical references are extracted from CNN news. The sound track is processed and the words overprinted on the image of each news story. An index is created relating locations with segments of news stories. This will allow, for instance, to query the system asking for news from a specific place, and to show a map of the location of the news story.

The interesting tool described by [Nobre et al. 2001] allows users to define the line of movement of the camera along a video over the space and also the region of space represented in each frame of the video. The result is that each frame is precisely geo-referenced. They used the tool to support a forest fire decision support system. A GPS is synchronized with the camera in order to store its position in every instant of the recording, which allows them to compute the position of each frame. When a fire starts, it is very easy to obtain video images of the area of the fire to help the firemen service.

**Description of our current VideoGIS system**

Our system has to deal both with a video base, and with geographic information. Section 3.1 discusses how are the videos in the video base, while section 3.2 focuses on the geographic information. Segmenting the videos using its geographic features and indexing the resulting segments videos are respectively described in section 3.3 and 3.4. Finally, how the querying mechanism works and how the results are shown are both presented in section 3.5.

**The video base**

The video base is built from a set of videos, concretely QuickTime movies. Although QuickTime has its own media file format, it does support a lot of different formats and codecs. Each described using an XML document. This document includes not only the source and frame rate of the video, but also the geographic coordinates of each frame in the video.

```xml
<video src="v0.mov" fps="15">
  <frame>
    <number>0</number>
    <x1>347003</x1>
    <y1>4700506</y1>
    <x2>347403</x2>
    <y2>4700806</y2>
  </frame>
  ...
</video>
```

*Figure 1. Example of a fragment of a video description XML document*

The video base is open and any user is allowed to insert new material, assuming s/he attaches the description document. A web-based tool will be developed to allow users to insert materials.
Further developments will provide a component that will use more easy-to-capture parameters to describe the video. They will be related to the camera, as its position, focal length and orientation (in the three axes). These three values make it possible to obtain the coordinates of the frame. This tool has to take into consideration that GPS receivers are not capable of storing positions each 1/24 seconds, and consequently the tool will interpolate between different measurements.

**Geographic information**

The geographic information consists of vector features which are imported into the system using GML 2.0 [OGC 2001]. The geometry types of these features can be the ones standardized by GML 2.0:

- Point
- LineString
- LinearRing
- Polygon
- MultiPoint
- MultiLineString
- MultiPolygon
- MultiGeometry

The system uses PostgreSQL, a RDBMS with spatial indexing capabilities by means of R-trees [Guttman 1984]. It imports spatial data from the GML documents and inserts the elements into the database, using one table for each type of features. It also updates two dictionary tables, one describing the tables of features and other describing the columns of each table. These tables will be used later to build dynamic queries on the features. This allows the system to be open to new data to be added. Since this process is automatic, an expert is not needed for this operation.

In future versions, the source of geographic information will be a set of geographical information systems and relational databases with vectorial information. OpenGIS specifications [OpenGIS specifications URL] will be used to integrate them.

**Video segmentation**

The first problem when dealing with video sources is how to segment it, i.e. how to choose the fragments of video that will be the base of later indexing and search. In the case of text, segmentation is quite simple, because the minimum unit of meaning is clear: the word; and it is trivial to extract the words of a text. But in the case of video, it is much more complex to know which are these units, especially from a semantic content point of view. The segmentation is closely related to the searches which are possible later.

One option is a handmade segmentation of video, but this is too expensive for big archives. Moreover, manual indexing has other problems as [Smeaton 2000] points out:

- No consistency of interpretation by a single person over time
- No consistency of interpretation among a population of interpreters
- No universally agreed format of the representation, whether keyword, captions or some knowledge-based formalism

Due to these reasons, automatic segmentation of video has been recently an intensive research field. Algorithms detect shots with a good reliability, but the parameters extracted from them and then used for indexing are of a very low level. These low level parameters are usually very basic geometric image properties and camera movements and have a poor semantic GI value. This semantics is needed to automatic production videos, in our case in response to two types of queries that the system must support:

- Retrieve all the features in one segment, needed to overprint the elements of interest on a segment
- Retrieve all the segment showing one (or a set of) feature(s), used to select the segments that will form part of a final video

The semantic information we want to attach to segments is related to the geographic features present in the video. Thus, we use both the coordinates and existing geographic features for automatic segmentation of videos. Consequently, a segment is defined as a set of contiguous frames containing the same set of features. The algorithm checks for every frame its set of features. This set will be compared with the set of the previous frame, and if they are different a new segment is started. Although currently the algorithm only checks the inclusion of the same elements, the criteria to compare those two sets may be more complex taking into account the relevance of each feature.
For every frame in the video Do
  Get the coordinates of the frame (a polygon)
  Get the set of features in that polygon
  If this set “differs” from the previous one Then
    Finish segment
    Start new segment
  End If
End For
Finish last segment

Figure 2. Algorithm for segmentation based on geographic features

This approach does not take care of layers (considering a layer for each type of feature) when segmentation. The use of different segmentations, one for each layer would improve the retrieving of segments showing only one feature or a set of features of a concrete layer. But it is not appropriate if working with several layers at the same time.

Indexing

We have used a variation of the Frame-segment tree [Subramanian 1997] to index and retrieve segments. The frame-segment tree combines a list of the features with a list of the segments of video. Each feature has a dynamic list with pointers to the segments where it is present, while each segment has a dynamic list with pointers to its features. These segments are indexed using a segment tree [Bentley 1977]. But since segments are stored in a PostgreSQL database, a typical B-tree is used to index them, instead of the segment tree proposed by Subramanian. On the other hand, the geographic features are indexed using R-trees.

This structure makes it possible an efficient retrieval of both the segments containing a given feature, and the features contained in a given segment.
Querying the system and showing the results

As it has been pointed out at section 3.3, the system has to support two types of queries, which goals are:

- Overprint the features and their information on the video when and where they are displayed
- Retrieve an on-the-fly created video according to a spatial query

The current system works with QuickTime for Java, the Java API for QuickTime. QuickTime offers a powerful architecture, the Media Abstraction Layer, which defines how software tools and applications can access the set of media support services built into QuickTime. The weakest point is that QuickTime for Java is in fact only a wrapper to the native C API of QuickTime, and therefore only works over Windows and Macintosh.

A movie is the object storing time-based data (audio and video) in QuickTime [Maremaa et al. 1999]. It can be self-contained or can handle references to other sources of data. It can contain more than one stream of data, each one called a track. The track does not contain data itself but references to media data. The player can show or hide these tracks. A special type of tracks, HREF tracks, allows hyper-linking. These hyperlinks can be automatic or may need user clicks to traverse them.

The video resulting of VideoGIS is a QuickTime movie which contains a track with the video information and one track for every layer of the geographical information. In that way, the user (or the system) can hide or layers (tracks) or show them as in a GIS. An HREF track is also provided to link to the information of features.

The system works in the form of a Java servlet. It generates the final video and the HTML page which inserts it, and they are sent to the client, which only needs a browser with QuickTime plug-in.

But the system is expected not to be closed to QuickTime or other architecture. A set of components will be built in future versions to allow different possibilities, in both the server and the client. Once the system knows the segments and features to retrieve, it will not generate a QuickTime video, but an XML document describing it. This XML document will be interpreted by different components supporting other types of presentations, such as the emerging standard SMIL [SMIL URL].

This approach gives the flexibility to allow the processing of the XML document in the server or in the client, while the transmission protocol will also be open to HTTP, RTP or other.

This pays a small cost of an overloading due to the extra step of generating and then parsing the XML document. But the advantages, in terms of flexibility and cross-platform, do worth. Other architectures will also be analysed. Nevertheless, the definitive architecture will also support different GI sources by means of OpenGIS specifications, as it was pointed out at section 3.2.

Future directions of work

Besides the definition of the final architecture, two important issues have to be addressed: video models and navigation.

Video models

The current system shows finally a video made of all the fragments satisfying a query, but the result may be unmanageable when the video base grows. It could be too long, with repeated zones, without any thread. More complex models of video are needed in order to build a video with a meaningful narrative thread. We have analyzed several semantic models, the most relevant are outlined here.

A first approach is [Bibiloni 1999] which uses a hierarchy of levels of segments aggregation. It also defines an algebra and an edition language to compose sequences, based on the semantic information associated to each segment. We call this a model based on themes, because typically the information of each segment corresponds has components from different themes. The aggregations of segments to compose sequences will differ if the focus is on one or other theme.

[Aguierre Smith et al. 1991] and [Davenport et al. 1991] propose a system where instead of partitioning the footage, the contextual information is segmented. The result is not a list of segments with a related semantic information, but the opposite: a list of descriptions each one associated to a starting and ending frame. This subtle difference allows them to define a model for automatic composition of scenes based on the film theory. However they do not implement the rules for composition, only provide the conceptual framework. This model is called the stratification model because each descriptive attribute is called a stratum, and it will be the base for the composition.

Following this approach, [Hjelsvold et al. 1994] proposes a more flexible model, where a video is composed by a hierarchy of structural components and each of these is associated to a sequence of frames. However, the
content is often associated to arbitrary sequences of frames, not related to the hierarchy of descriptions that has been created. Their model supports the definition of annotations independent from the structure for that content.

A different approach from [Tran et al. 2000] supports the definition of semantic content without any temporal value and build compositions using the thirteen temporal relations from Allen, it is possible to build compositions.

Currently we are working in the definition of a new semantic model which also takes into account the special properties of the geographic information of VideoGIS. A first possibility is to infer a thematic information from the set of geographic elements of a segment in order to use later a thematic model similar to [Bibiloni 1999].

Navigation

Another key issue is to improve navigation. One of our main goals is to allow browsing of videos in a spatial way. However, the current system is focused on retrieval and is quite poor on navigation. Although the video is dynamically generated and breaks the linearity of the sources, the user can only play it in a sequential way.

A first idea is to consider intersections among videos. This will allow the user to move from one video to another making a sort of virtual environment. This idea is in fact not new and was already implemented in the Aspen Movie Map project, and it is typical in video games, but has been always applied to a restricted set of videos. In VideoGIS, the system would show to the user what s/he will find following each choice. But it could be more useful adding personalization capabilities, allowing the system itself to decide which is the best choice according to the user preferences. An interesting reference about video browsing is [Li et al. 2000] where the authors study it on different content areas (such as entertainment, news, ...) and propose different behavioural patterns of the experimental work they carried out. Most of the conclusions are about the enhanced controls allowed by digital video.

The definition of the new semantic model of video will require a deeper analysis of new possibilities of thematic navigation. It is likely that a methodology for creating navigation from models, such as the multimedia methodologies analysed in [Navarrete 1998], will be proposed, taking into account recent alternatives to the subject.

Conclusions

We have presented a novel approach to combine video and geographic information. The system uses automatic segmentation based on coordinates and geographic features as well as an indexing structure, which is a variation of the frame-segment tree. It generates videomaps i.e. videos with superimposed information of geographic features, creating a map where a video is on the bottom layer. It is also able to generate videoitineraries in response to spatial queries, i.e. a composition of sequences showing the geographic features that satisfy the query. We propose videomaps and videoitineraries as an alternative interface to traditional maps for non-skilled users.

We have discussed the need for a semantic model that will allow to compose videoitineraries with a more meaningful narrative thread and the need of improving the poor navigability of the current approach.

References


Hjelsvold, R.; Midtstrøm, R., 1994: Modelling and querying data. Proceeding of *20th International Conference on Very Large Data Bases*, Santiago (Chile).


OpenGIS Specifications URL: [http://www.opengeos.org/ogcSpecs.htm](http://www.opengeos.org/ogcSpecs.htm)


PostgreSQL URL: [http://www.postgresql.org/](http://www.postgresql.org/)


Smeaton, A.F., 2000: Indexing, Browsing and Searching Digital Video and Digital Audio Information. Presented at the *3rd European Summer School on Information Retrieval*, Varenna (Italy).

SMIL URL: [http://www.w3.org/AudioVideo/](http://www.w3.org/AudioVideo/)


XML URL: [http://www.w3.org/XML/](http://www.w3.org/XML/)