

Mining Repeated Patterns in Broadcast Multimedia

In the morning of September 11, 2001, two French filmmakers were shooting a documentary about New York firefighters at an intersection less than a mile north of the World Trade Center. While they were interviewing a group of firefighters, they heard a roar from above and turned the camera to capture the only known video of the first plane crashing into the north tower of the World Trade Center [1]. Almost immediately, this short video footage was aired in all the major news channels, marking the beginning the worst ever foreign attack on the American soil.

On September 21, 1998, the U.S. House Judiciary Committee released President Bill Clinton's videotaped testimony in the Monica Lewinsky affair. Arguably the most popular video footage that year, President Clinton's testimony was repeatedly broadcasted, either in its entirety or highlights, by television stations worldwide. The affair with Monica Lewinsky brought President Clinton to the edge of impeachment, nearly costing his presidency.

Radio and television broadcasting has been a major influence in shaping the political, social, cultural, and economic trends of the twentieth century. Historical events, like the two mentioned above, are usually marked by repeated airings of the same video clips or sound-bytes. How many times have we seen the footage of the falling twin towers, President Kennedy's assassination, or Dr. Martin Luther King's speech in Washington D.C.? The broadcast frequency of a particular video or audio clip, especially over a long period of time and in diverse geographical locations, is perhaps one of the most important indicators of its historical significance. It has long been established that a key selection criteria for long-term archival of television broadcasting is "material of historical interest in all fields." [2]. As a result, the capability of finding the frequency of repeated broadcast of the same video or audio material should greatly facilitate the automatic ingestion and preservation of historically important content.

Not all repeated broadcast, however, are historically significant. Perhaps the largest category of repeated broadcast belongs to television commercials. Even though finding repeated commercials may not be important to historians, it is crucial for advertisers to track the air time of their commercials and for companies to monitor new products from their competitors. In 1997, a scandal broke out in Japan when advertisers found out that they were paying for thousands of commercials that were never aired [3]. Such a practice of overbooking the airtime remained undetected for more than 20 years because there was no system in place for monitoring broadcast advertisements.

While large corporations may hire specialized firms to track their own commercials, there are situations where broadcast monitoring needs to be performed as a public service by an independent organization. One such example is the monitoring of political campaign advertisements. The Bipartisan Campaign Reform Act of 2002 states that it is a criminal offense for any corporations or labor unions to sponsor television advertisements for a candidate 60 days before a general election or 30 days of a primary [6]. An automatic system that can find not only the frequency but also the time and station of each airing of a political advertisement will be indispensable for enforcing such a law.

The goal of this project is to develop reliable and efficient techniques to identify all repeated patterns from a large number broadcast television and radio channels for a prolong period of time.

We believe that detecting repeated patterns is an important tool for long-term television archiving in identifying historically important content and monitoring broadcast commercials. In the initial phase of the project, we intend to build a prototype system that is capable of monitoring roughly 70 local and national television channels, as well as 120 radio channels from satellite radio. In the future, we plan to collaborate with archivists in better understanding the captured repeated contents, to develop automatic techniques for classifying them into different genres, and to expand our monitoring to internet broadcast.

It is important to realize that the easily-available television and radio programming guides do not provide sufficient granularity to mine most repeated patterns of interest. Newly developed multimedia meta-data standards such MPEG-7 does provide a framework for describing multimedia content in minute details [5]. Nonetheless, they are not widely used because of the high cost in generating and embedding the meta-data with the media as well as the administrative barrier of using an universal identifier for every multimedia item. Thus, we argue that the only viable approach available today to identify a multimedia item is by extracting salient audiovisual features directly from the content. A content-based approach requires us to process an enormous amount of information. Assuming the output data rates for a typical television station and a radio station are 0.05 GB/hour and 2.25 GB/hour [4]. One year worth of broadcast from 70 television and 120 radio stations amounts to 1,347 terabytes. Mining repeated patterns from such a large streaming database pose many technical challenges that cannot be met by existing research in content-based multimedia recognition and data mining. In this project, we plan to develop highly discriminatory and low-complexity audiovisual features to capture the essence of the content, and invent new summarization and data-mining algorithms that can robustly identify arbitrary-length repeated occurrences of these features.

References

- [1] Cultural Services of the French Embassy in the U.S. <http://www.info-france-usa.org/culture/tv/programs/naudet911.html>, New York, N.Y. *9/11 A Documentary by Jules and Gedeon Naudet*.
- [2] M. Ide, D. MacCarn, T. Shepard, and L. Weisse. Understanding the preservation challenge of digital television. In *Preserving our digital heritage: Plan for the National Digital Information Infrastructure and Preservation Program. A Collaborative Initiative of the Library of Congress*, chapter Appendix 2. Library of Congress, 2002.
- [3] D. Kilburn. Dirty linen, dark secrets. *Adweek*, 38(40):35–40, October 1997.
- [4] P. Lyman and H. R. Varian. *How much information 2003?* School of Information Management and Systems, University of California at Berkeley, <http://www.sims.berkeley.edu/research/projects/how-much-info-2003/index.htm>.
- [5] B.S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Ltd., 2002.
- [6] U.S. Government. *The Bipartisan Campaign Reform Act of 2002*, public law no. 107-155 edition, March 2002.