# TalkMiner: A Search Engine for Online Lecture Video

John Adcock[1] Matthew Cooper[1] Laurent Denoue[1] Hamed Pirsiavash[2]
Lawrence A. Rowe[1]

[1]FX Palo Alto Laboratory
3400 Hillview Ave.
Palo Alto, CA 94304 USA
{last name}@fxpal.com

[2]Dept. of Computer Science
University of California, Irvine
Irvine, CA 92697 USA
hpirsiav@ics.uci.edu

## ABSTRACT

TalkMiner is a search engine for lecture webcasts. Lecture videos are processed to recover a set of distinct slide images and OCR is used to generate a list of indexable terms from the slides. On our prototype system, users can search and browse lists of lectures, slides in a specific lecture, and play the lecture video. Over 10,000 lecture videos have been indexed from a variety of sources. A public website now allows users to experiment with the search engine.

## Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

## General Terms

Algorithms, Experimentation

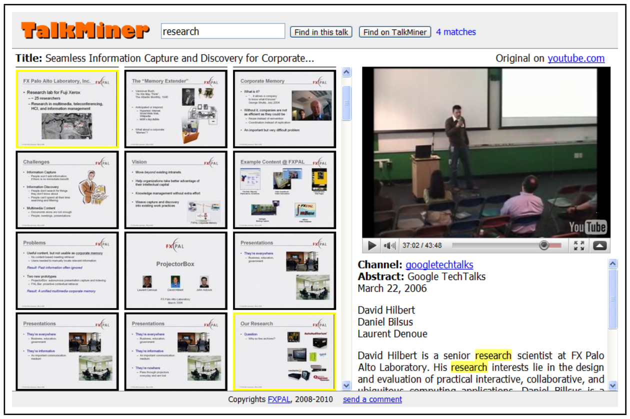
## 1 Overview

Lecture webcasts are readily available on the Internet. These webcasts include class lectures (e.g., Berkeley Webcast, MIT Open Courseware, etc.), research seminars (e.g., Google Tech Talks, PARC Forums, etc.), product demonstrations, or training materials.

Conventional web search engines can be used to locate these lecture videos, but only when supporting text appears on the hosting web page, or the media has been tagged or otherwise authored within a purposed hosting system. But users, especially students, need to find the locations *within* a video when an instructor discusses a specific topic. Addressing this need requires a search engine that can identify relevant material with the content of the webcast.

TalkMiner provides an enhanced search and browsing system designed to improve the usefulness of lecture webcasts by enabling keyword search of slides appearing within a presentation video, and video playback directly from the time a chosen slide appears (see Figure 1). The system analyzes the videos to identify and extract unique slide images along with their time codes. This allows seeking the embedded video to the appearance of each detected slide. OCR is applied to detected slide images and a text search index is built from the recovered words.

TalkMiner builds its search index and interface from commonly recorded video. It requires neither dedicated lecture-capture systems, nor careful post-capture authoring, nor even constraints on the style of the video capture. Thus, the system can scale to include a greater volume and variety of existing and newly created content at a much lower cost than would otherwise be possible. Also, by leveraging existing online video distribution infrastructure to embed webcasts within an enhanced interface the system minimizes storage and bandwidth requirements, further aiding scalability and portability.



**Title:** Seamless Information Capture and Discovery for Corporate...

Original on youtube.com

**Channel:** googletechtalks
**Abstract:** Google TechTalks
March 22, 2006

David Hilbert
Daniel Bilsus
Laurent Denoue

David Hilbert is a senior research scientist at FX Palo Alto Laboratory. His research interests lie in the design and evaluation of practical interactive, collaborative, and

Copyrights FXPAL, 2008-2010 send a comment

**Figure 1: Viewing slides and using them to seek the embedded player. Slides matching the user-specified search terms are highlighted.**

The system currently indexes lecture videos from three sites: YouTube [7], U.C. Berkeley [1], and blip.tv [2]. The current number of talks indexed is over 11,000.

## 2 System Architecture

An overview of the architecture appears in Figure 2. The system is composed of two main components: the back-end video indexer and the front-end web server. The video indexer searches the web for lecture webcasts and indexes them. It automatically identifies slide images and processes them to create the text search index.

The method for identifying and accessing the presentation media varies slightly for each source site, but generally videos are identified by parsing RSS feeds to which the system is subscribed. Once downloaded, it takes roughly 6 minutes to process a 60 minute talk (i.e., 10 percent of the real time talk duration), so the system is generally limited by download speed rather than processing speed.

The TalkMiner web-based front end is implemented in Java server pages (JSP) running on an industry standard

*MM'10,* October 25–29, 2010, Firenze, Italy.
ACM 978-1-60558-933-6/10/10.

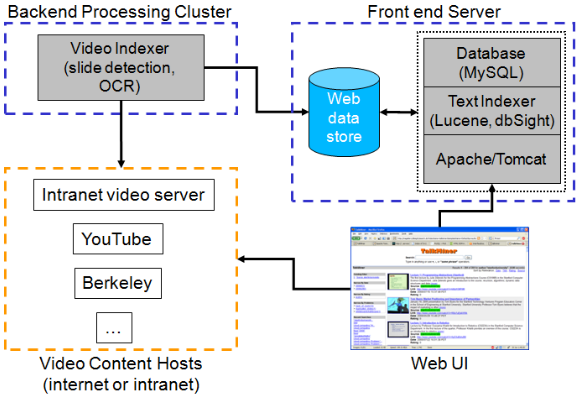

Backend Processing Cluster
Video Indexer (slide detection, OCR)
Front end Server
Web data store
Database (MySQL)
Text Indexer (Lucene, dbSight)
Apache/Tomcat
Intranet video server
YouTube
Berkeley
...
Video Content Hosts (internet or intranet)
Web UI

**Figure 2: System architecture showing separate back-end, front-end, and content hosting components.**

Apache/Tomcat combination. The indexing and search framework, implemented with DBSight [4], runs in the same Tomcat instance. At runtime, searches are performed through the DBSight web application from previously computed Lucene indexes of the MySQL talk database to render the search results lists. The detailed talk page reads the detailed slide text and timing information directly from the MySQL database. Embedded player control and within-talk text searches are performed in javascript, freeing the system from futher web or database access until the next search.

## 3 User Interaction

The TalkMiner search interface resembles typical web search interfaces. The user enters one or more search terms and a list of talks that include those terms in the title, abstract or the presentation slides are listed as shown in Figure 3.

The information displayed for each talk on the search results pages includes a representative key frame, the title of the lecture, and the channel or source of the talk. Other metadata displayed includes the duration of the talk, the number of slides, the publication date, and the date it was indexed by TalkMiner. An attribution and link to the original video source is also provided.

The user can browse a list of talks and alter the sorting and filtering criteria for the listing. By default, talks are sorted by relevance to the query terms as computed by the Lucene text retrieval software [6]. Other available sort attributes include publication date, number of slides, channel, and rating. The first column on the left side of the results page includes interface controls to filter results according to specific criteria (e.g., the year of publication, the channel, etc.). It also includes a list of recent search queries to allow users to re-execute a recent query.

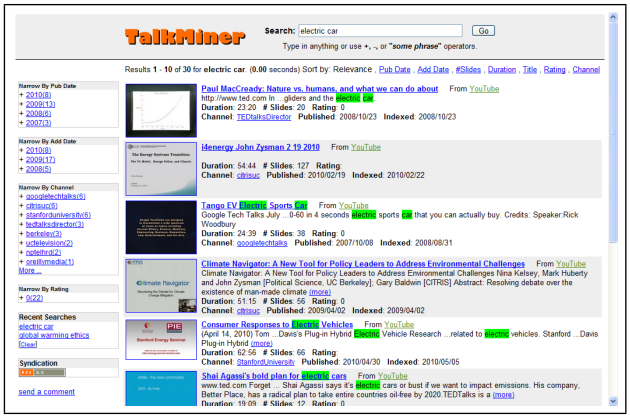
Search results link to the detailed talk view as depicted in Figure 1. Slides matching the query are highlighted, and the user can control the playback position of the embedded video player by selecting the slide thumbnail with the content of interest.

## 4 Slide Detection

We adapted a straightforward frame-difference based analysis from ProjectorBox [5] to identify keyframes for TalkMiner. We extract a keyframe for nearly stationary segments of a minimum length; these generally correspond to slides. We have extended this baseline approach to address several frequently occurring cases in which it often selects spurious keyframes, or misses slide appearances altogether.

- full-frame shots of the speaker with neither navigational or informational value.
- shots of slides that contain “picture-in-picture” streams.
- shots from the back of the room that include the audience in the foreground and/or the speaker.

To improve our slide detection we have incorporated spatial filtering to ameliorate the effects of insignificant motion, and self-bootstrapping visual models to filter out shots of the speaker.



electric car

Type in anything or use +, -, or "*some phrase*" operators.

Results **1** - **10** of **30** for **electric car**. (**0.00** seconds) Sort by: Relevance , Pub Date , Add Date , #Slides , Duration , Title , Rating , Channel

**Narrow By Pub Date**
+ 2010(8)
+ 2009(13)
+ 2008(6)
+ 2007(3)

**Narrow By Add Date**
+ 2010(8)
+ 2009(17)
+ 2008(5)

**Narrow By Channel**
+ googletechtalks(6)
+ citrisuc(6)
+ stanforduniversity(6)
+ tedtalksdirector(3)
+ berkeley(3)
+ uctelevision(2)
+ nptelhrd(2)
+ oreillymedia(1)
More ...

**Narrow By Rating**
+ 0(22)

**Recent Searches**
electric car
global warming ethics
[Clear]

**Syndication**
RSS 2.0

send a comment

Paul MacCready: Nature vs. humans, and what we can do about From YouTube
http://www.ted.com In ...gliders and the electric car.
**Duration**: 23:20 **# Slides**: 20 **Rating**: 0
**Channel**: TEDtalksDirector **Published**: 2008/10/23 **Indexed**: 2008/10/23

i4energy John Zysman 2 19 2010 From YouTube
**Duration**: 54:44 **# Slides**: 127 **Rating**:
**Channel**: citrisuc **Published**: 2010/02/19 **Indexed**: 2010/02/22

Tango EV Electric Sports Car From YouTube
Google Tech Talks July ...0-60 in 4 seconds electric sports car that you can actually buy. Credits: Speaker:Rick Woodbury
**Duration**: 24:39 **# Slides**: 38 **Rating**: 0
**Channel**: googletechtalks **Published**: 2007/10/08 **Indexed**: 2008/08/31

Climate Navigator: A New Tool for Policy Leaders to Address Environmental Challenges From YouTube
Climate Navigator: A New Tool for Policy Leaders to Address Environmental Challenges Nina Kelsey, Mark Huberty and John Zysman [Political Science, UC Berkeley]; Gary Baldwin [CITRIS] Abstract: Resolving debate over the existence of man-made climate (more)
**Duration**: 51:15 **# Slides**: 56 **Rating**: 0
**Channel**: citrisuc **Published**: 2009/04/02 **Indexed**: 2009/04/02

Consumer Responses to Electric Vehicles From YouTube
(April 14, 2010) Tom ...Davis's Plug-in Hybrid Electric Vehicle Research ...related to electric vehicles. Stanford ...Davis Plug-in Hybrid (more)
**Duration**: 62:56 **# Slides**: 66 **Rating**:
**Channel**: StanfordUniversity **Published**: 2010/04/30 **Indexed**: 2010/05/05

Shai Agassi's bold plan for electric cars From YouTube
www.ted.com Forget ... Shai Agassi says it's electric cars or bust if we want to impact emissions. His company, Better Place, has a radical plan to take entire countries oil-free by 2020.TEDTalks is a (more)
**Duration**: 19:09 **# Slides**: 12 **Rating**: 0

**Figure 3: Search results. Each lecture shows a representative keyframe, attribution, title, and description when available. Results can be filtered or sorted based on metadata including the date, duration, number of slides, and source of the webcast.**

## 5 Copyright Considerations

Video on the web exists under a wide variety of copyrights and terms of use and the implementation of TalkMiner has taken this into account. University lecture material usually has an explicit Creative Commons [3] license, but even these vary in their scope, in particular not always allowing for the creation of derivative works which puts various desirable modifications on shaky legal ground. The potential value of TalkMiner is much higher when using content without copyright restrictions, such as would be the case if the system were deployed by the copyright holder proper.

## 6 References

[1] Berkeley Webcasts. http://webcast.berkeley.edu/.

[2] Blip TV. http://www.blip.tv/.

[3] Creative Commons. http://creativecommons.org, 2007.

[4] DBSight. http://www.dbsight.net/.

[5] L. Denoue, D. Hilbert, D. Billsus, and M. Cooper. Projectorbox: Seamless presentation capture for classrooms. In *World Conf. on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 2005.

[6] Apache Lucene. http://lucene.apache.org/java/docs/.

[7] YouTube. http://www.youtube.com/.