Columbia University's Semantic Video Search Engine

Shih-Fu Chang, Lyndon S. Kennedy, Eric Zavesky[1]

ABSTRACT

We briefly describe "CuVid," Columbia University's video search engine, a system that enables semantic multimodal search over video broadcast news collections. The system was developed and first evaluated for the NIST TRECVID 2005 benchmark and later expanded to include a large number (374) of visual concept detectors. Our focus is on comparative studies of pros and cons of search methods built on various individual modalities (keyword, image, near-duplicate, and semantic concept) and combinations, without requiring advanced tools and interfaces for interactive search.

Categories and Subject Descriptors: H.3.3 **[Information Storage and Retrieval]**: Information Search and Retrieval.

General Terms: Algorithms, Design, Experimentation

Keywords: Video Search & Retrieval, VideOlympics Showcase

1. INTRODUCTION

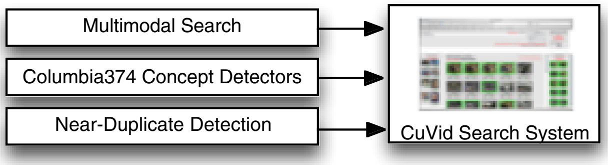
The Columbia video search system provides a platform for users to semantically search and browse broadcast news video corpora through a simple interface which can be accessed with any modern web browser. The system incorporates many key components, such as text search with story segmentation [2], visual example-based search with near-duplicate pairs [3], and concept-based search with a large set of pre-trained detectors [4]. Each component is briefly described below and discussed and evaluated in greater detail in [1].

2. SYSTEM COMPONENTS

2.1 Concept-based Search

The use of pre-trained visual concept detectors has been shown to be a powerful tool for generating and filtering results in video search systems. At the core of the Columbia system, is a very large lexicon of 374 concept detectors [4], which are applied to each subshot in the collection, giving a semantic representation of the content of the subshot. Users of the system can interact with the detectors on a number of levels. An advanced user might construct a query with a combination of any number of detector results, such as using the "building" and "protest" detectors to find scenes with both buildings and protests present. A novice user might struggle when dealing with such a large collection of concepts and instead issue a query of text keywords and have the system automatically map from this textual input to a combination of concepts.

2.2 Text Search



Search over text transcripts (from speech recognition and machine translation) is an essential component of any video search system, particularly for the broadcast news domain, where shots of named persons can reliably be found by simply looking at shots adjacent to the mention of the persons name in the speech transcript. An important issue with the use of text search is the asynchrony between the times at which words are spoken and the times at which visual concepts appear in the video stream. To deal with this issue, we group text and shots together into stories, which are segments with a single, coherent news focus. Story boundaries are automatically detected [2] and text search is conducted using stories as documents. The text retrieval score for a story is propagated to all shots occurring temporally within the story boundaries.

2.3 Visual Example-based Search

Matching shots in a search set to example shots provided by a user is a common approach to searching over visual content, which can be very powerful in interactive retrieval systems, where users can iteratively provide relevance feedback. We incorporate two types of visual matching in our system. The first is a generic matching based on low-level features such as colors in a 5x5 grid and edges and texture features over the whole image. The second is a highly powerful detector of *near-duplicate* pairs [3], which are defined as images of the same scene which are nearly identical, but may differ due to the use of two different cameras or other post-processing effects. These near-duplicate pairs have been shown to be very powerful for search over parallel news corpora (such as TRECVID 2005) [1], since many such duplicates appear due to the effect of the same news story being covered by many different news broadcasts.

3. REFERENCES

[1] S.-F. Chang *et al*. Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction. *NIST TRECVID Workshop,* Gaithersburg, Nov. 2005.

[2] W. Hsu, S.-F. Chang. Visual Cue Cluster Construction via Information Bottleneck Principle and Kernel Density Estimation. In *International Conference on Image and Video Retrieval (CIVR)*, Singapore, July 2005.

[3] D.-Q. Zhang, S.-F. Chang. Detecting Image Near-Duplicate by Stochastic Attributed Relational Graph Matching with Learning. In *ACM Multimedia*, New York, Oct. 2004.

[4] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. Columbia ADVENT Technical Report #222-2006-8, March 2007. Download site: http://www.ee.columbia.edu/dvmm/columbia374.

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands.
Copyright 2007 ACM 978-1-59593-733-9/07/0007.

[1] Dept. of Electrical Engineering, Columbia University, New York, NY 10027, {sfchang,lyndon,emz}@ee.columbia.edu