The AXES PRO Video Search System

Kevin McGuinness,
Noel E. O'Connor
CLARITY: Centre for Sensor Web Technologies
Dublin City University, Ireland

Robin Aly,
Franciska De Jong
University Twente,
Netherlands

Ken Chatfield,
Omkar M. Parkhi,
Relja Arandjelovic,
Andrew Zisserman
University of Oxford, UK

Matthijs Douze,
Cordelia Schmid
INRIA, France

ABSTRACT

We demonstrate a multimedia content information retrieval engine developed for audiovisual digital libraries targeted at media professionals. It is the first of three multimedia IR systems being developed by the AXES project. The system brings together traditional text IR and state-of-the-art content indexing and retrieval technologies to allow users to search and browse digital libraries in novel ways. Key features include: metadata and ASR search and filtering, on-the-fly visual concept classification (categories, faces, places, and logos), and similarity search (instances and faces).

Categories and Subject Descriptors

I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision—*Applications*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Video*

General Terms

Design, Experimentation, Human Factor, Algorithms

Keywords

Multimedia IR, Computer Vision, Video

1. INTRODUCTION

AXES is an EU FP7 project aimed at developing tools that provide various types of users with new engaging ways to interact with audiovisual libraries, helping them discover, browse, navigate, search, and enrich archives. To achieve this goal, the project is developing a series of digital library search and navigation systems tailored for different user groups: professional users, researchers, and home users. The AXES PRO system that we will demonstrate targets the first of these groups: the professional user. The system brings together traditional text based IR techniques and state-of-the-art computer vision and content based multimedia search technologies, enabling the end user to leverage this combination of technologies in novel ways.

ICMR'13, April 16–20, 2013, Dallas, Texas, USA.
ACM X-XXXXX-XX-X/XX/XX.

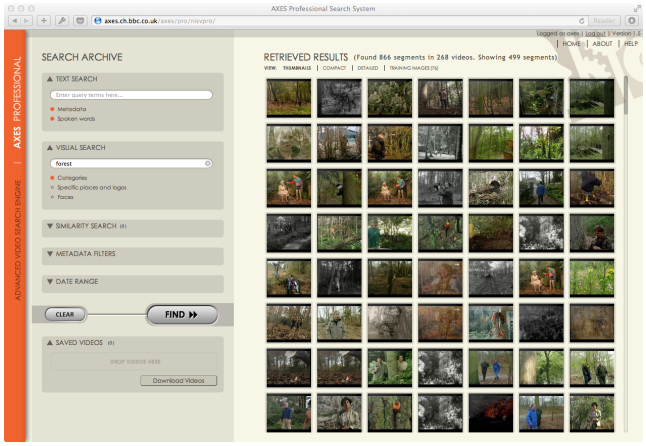

Figure 1: Screenshot of the AXES PRO system showing the thumbnail view of the search results.

2. SYSTEM OVERVIEW

User interface. AXES PRO has a browser-based UI composed of two panels: search archive and retrieved results. The search archive panel allows users to formulate queries using text and images. It supports visual search, metadata and ASR text search, metadata and date range filters, visual similarity search, and video saving and download. The results panel shows the query results in various ways.

The user interface includes three different result views: thumbnails, compact, and detailed. The thumbnails view (Figure 1) shows only thumbnails of the key frames associated with each result, allowing the user to visually browse through many results quickly. This view is most useful when the user is primarily interested in the visual appearance of results, as may be the case in a known-item or instance search scenario. The detailed view (Figure 2) shows more detailed information about each result, including the video title, language, creator/publisher, creation date, license, video description, clip duration, position of the clip in the video, and information about why the clip was retrieved (matched on text, or visual similarity, etc.). This view is most useful in scenarios where the user wishes to retrieve video segments by title or description. The extra detail, however, means that

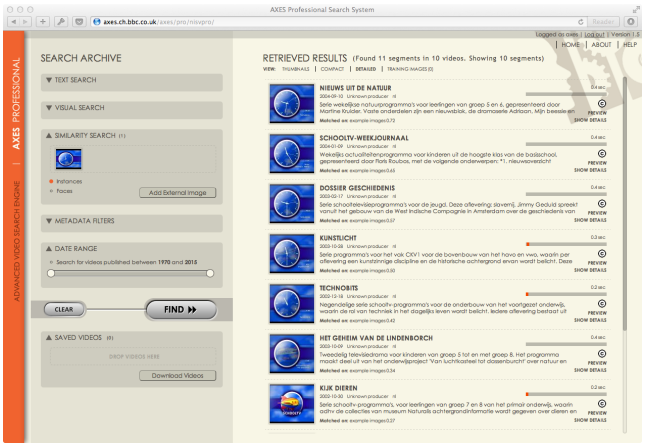

Figure 2: The detailed view of the search results.

fewer results can be shown simultaneously to the user. The compact view provides a middle ground between the detailed and thumbnail view, showing less information in a compact layout that allows more results to be shown simultaneously.

Text search. The system stores and indexes all metadata and spoken words available at index time. Spoken words are extracted from content using automatic speech recognition, but can also be provided in the form of transcripts. The user interface supports basic text-based search of these fields using a standard search and check box interface. This allows, for example, the user to search for videos by title, by description, or for videos containing specific spoken phrases. Queries may also use standard IR Boolean conjunctives, such as AND and OR. Users can also filter results on specific metadata fields and constrain the results by publication date.

Visual search. The user interface supports text-based queries that are used in conjunction with an external search engine to collect exemplars and train visual concept classifiers on-the-fly. The text query from the user is used to gather a representative sample of images from an external source. The current implementation uses Google Images to find the top-n images for the query. The system also retains a fixed collection of arbitrary images assumed to be non-relevant. Using this set of positive and negative examples, the system then trains a discriminative classifier (for example, a support vector machine) using image descriptors extracted from the examples. The trained classifier is applied to each image in the dataset in turn to produce a score, and the result list created by ranking the dataset by score.

The system supports three types of visual search: visual categories, faces, and specific places or logos. By allowing the user to specify the type of search, the system can use features and classifiers tuned to that particular class of visual search. For example, when the user chooses visual search by faces, the system detects faces, locates facial features using a pictorial structures based method, and extracts local descriptors at the detected facial landmarks. Technical details on the specific approaches and descriptors used can be found in [1, 2, 3, 4].

Similarity search. The system allows the user to drag and drop key frames from the retrieved results to be used for similarity search. Visual queries can consist of a single image, or multiple examples. The user can also add images from their own computer or from an external URL. Region of interest selection is also supported.

Like visual search by text, the similarity search supports user selectable search types. The currently supported options are instance search and face search. Instance based similarity search uses the BigImbaz engine described in [5]. Face similarity search uses a system based on facial landmarks. A set of 9 facial landmark points are detected, located on the eyes, nose, and mouth. The face image is then warped using a similarity transform so that the landmark points are mapped as close as possible to a canonical configuration of a frontal pose. For each of the 9 landmarks, a histograms of oriented gradients (HOG) descriptor is extracted and these are concatenated to form the face signature. The high-dimensional face signature is then compressed into a lower dimensional signature by means of a linear projection. The projection matrix has been obtained by an off-line metric learning algorithm [6] so that the L2 distance between signatures after projection is small for face signatures of the same person and large for signatures of different people. The compressed face signature is then matched to face signatures in the database to find similar faces across other videos.

3. DEMONSTRATION

The demonstration will show the live system running on 400 hours of video content from the Dutch broadcaster NISV. We will demonstrate using various queries the available search modalities, including: on-the-fly visual search for categories, faces, specific places and logos; similarity search on instances and faces, text based search on ASR and metadata, and metadata filtering of results.

Acknowledgments. This work is supported by the EU Project FP7 AXES ICT-269980.

4. REFERENCES

[1] R. Arandjelovic and A. Zisserman. Multiple queries for large scale specific object retrieval. In *Proceedings of the British Machine Vision Conference*, 2012.

[2] K. Chatfield and A. Zisserman. VISOR: Towards on-the-fly large-scale object category retrieval. In *Proceedings of ACCV*, 2012.

[3] O. M. Parkhi, A. Vedaldi, and A. Zisserman. On-the-fly specific person retrieval. In *Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services*, 2012.

[4] R. Aly, K. McGuinness, S. Chen, N. E. O'Connor, K. Chatfield, O. M. Parkhi, R. Arandjelovic, A. Zisserman, UK B. Fernando, T. Tuytelaars, J. Schwenninger, D. Oneata, M. Douze, J. Revaud, D. Potapov, H. Wang, Z. Harchaoui, J. Verbeek, and C. Schmid. AXES at TRECVID 2012. In *Proceedings of the TRECVid Workshop*, 2012.

[5] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010.

[6] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *Proceedings of ICCV*, 2009.