THE AXES RESEARCH VIDEO SEARCH SYSTEM

Kevin McGuinness[1], *Robin Aly*[2], *Ken Chatfield*[3], *Omkar M. Parkhi*[3], *Relja Arandjelovic*[3], *Matthijs Douze*[4], *Max Kemman*[5], *Martijn Kleppe*[5], *Peggy van der Kreeft*[6], *Kay Macquarrie*[6], *Alexey Ozerov*[7], *Noel E. O'Connor*[1], *Franciska De Jong*[2], *Andrew Zisserman*[3], *Cordelia Schmid*[4], *Patrick Perez*[7]

[1]Dublin City University, Ireland [2]University Twente, Netherlands [3]University of Oxford, UK
[4]INRIA, France [5]Erasmus University Rotterdam, Netherlands [6]Deutsche Welle, Germany [7]Technicolor, France

ABSTRACT

We will demonstrate a multimedia content information retrieval engine developed for audiovisual digital libraries targeted at academic researchers and journalists. It is the second of three multimedia IR systems being developed by the AXES project[1]. 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).

1. INTRODUCTION

The goal of the AXES project is to develop 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. This is approached from three perspectives (or axes): users, content, and technology.

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 Research system that we will demonstrate targets the second of these groups: academic researchers and journalists. 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.

AXES Research updates our AXES system for professional users (AXES PRO [1]) in several ways. The individual technologies being used for similarity search and on the fly concept classification have been updated to improve performance. The system now also supports automatic selection of high quality pre-trained classifiers based on user input. The user interface is completely new, and designed specifically to support the requirements of the research user group. It has many new features that differentiate it from our previous system for media professionals:

[1]http://www.axes-project.eu

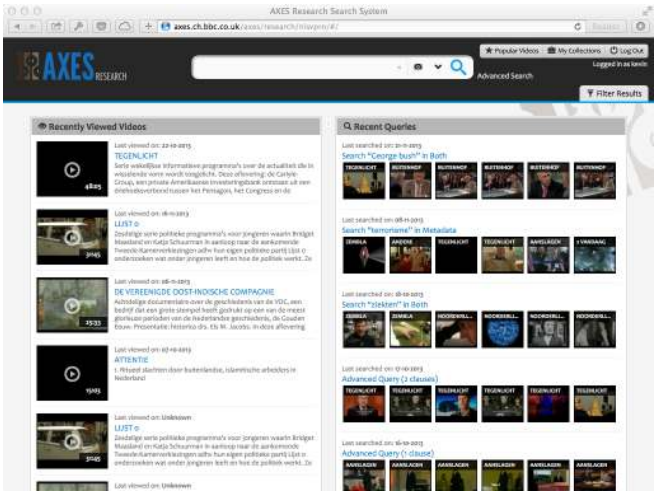

Fig. 1. The starting page for a user in the AXES Research interface provides users with a simple search interface and shows their recently viewed videos and queries.

Browsing and linking The AXES Research interface places a heavier emphasis on linking and browsing in addition to search. This is realized, for example, by having suggested video displayed along with individual results from a search, and by having known entities linked to queries.

Tagging and annotation Users can add their own annotations to documents in the form of public and private notes.

Search trails Retrieved documents may be the outcome of complex queries. These queries are stored and can be retrieved using deep links, so the user can always restore query results, bookmark results, and share results with colleagues.

Metadata export As metadata is important for academic users, journalists, and other research users, the research interface allows users to export metadata in several formats so that it can be collected and analysed by external tools.

Customization and personalization The research interface allows more customization and personalization than the AXES professional interface. The initial screen, for example, displays recent user searches and a list of recently viewed results. Also included is the capacity for the user to customize which features are enabled in the interface.

Collaboration The capacity to add public notes to search results can allow for collaborative research.

Usage statistics The new interface displays more information about the user's interaction with various media items. For example, the result list now displays information about the number of times a video was viewed, and the number of times it was downloaded.

2. SYSTEM OVERVIEW

The AXES Research interface is a browser-based, deep-linked, single page web application (SPA). The user interface communicates with an REST-like web service using AJAX and JSON. A centralized link management and structured search system is responsible for maintaining links, text indexes, and carrying out search and result fusion. The remainder of the search components, such as similarity search and visual concept classification, are generally distributed across multiple physical servers and communicate with the link management system using thin web-service interfaces. The following describes the major features in more detail.

Text search The system stores and indexes all metadata and spoken words available at index time. Spoken words are either extracted from content using ASR or provided in the form of transcripts. The UI supports simple text-based search of these fields using a standard search 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 system supports text-based queries that are used in conjunction with an external search engine to collect examples 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 the top-n results from Google Images. The system also retains a fixed collection of arbitrary images assumed to be non-relevant. Using these positive and negative examples, the system trains a discriminative classifier using image descriptors extracted from the examples. The trained classifier is applied to each image in the dataset to produce a score, and the result list is the dataset ranked by this 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 [2, 3, 4, 5].

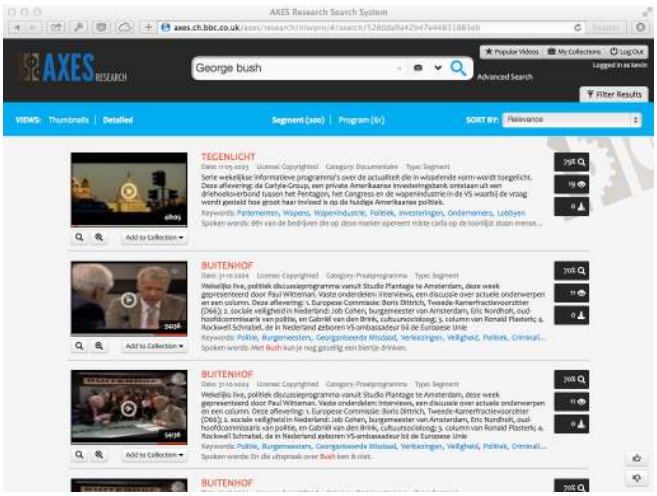

Fig. 2. The detailed results view shows the search results. Results can also be viewed as thumbnails.

Visual search currently takes approximately 20-30 seconds to complete, which includes time downloading example images, extracting features, training, and ranking. The system caches previous queries so that for common queries the results are almost instantaneous.

Similarity search Similarity search is supported both for in collection queries (using keyframes from returned results), and external images. Visual queries can consist of a single image or multiple examples. 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 [6]. Face similarity search uses a system based on facial landmarks. A set of 9 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 [7] 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.

Advanced Search One of the issues with the AXES PRO system was that, although it allowed the user to specify sophisticated queries, many found the query interface complicated. In AXES Research, we simplified the interface to a single

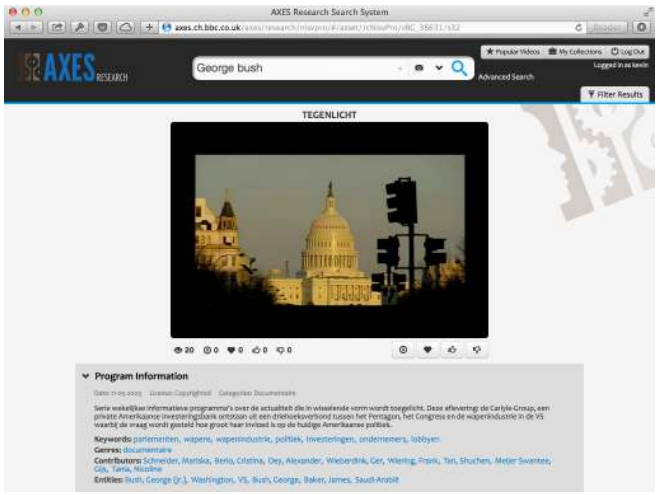

Fig. 3. The asset view showing a single video or segment and associated metadata. Audio transcripts, user notes, and related videos are displayed further down the page.

search bar with drop down options. This simplicity, however, comes at the expense of expressiveness: it is no longer possible to specify queries that fuse the outputs of, for example, an on the fly visual classifier and a text search on the output of automatic speech recognition. The advanced search interface caters for such queries without adding additional complexity to the standard search interface. Advanced searches are specified in terms of a set of text and similarity search clauses, each clause targeting a different search modality (e.g. visual categories, or similar faces). The backend forwards each clause to the relevant component and fuses the results.

Interacting with results Users can interact with videos and segments in many ways. They can preview the video in a popup player, or drill down for more detailed information. Users can playback videos, browse metadata and transcripts, and browse a video by keyframe. A virtual cutter allows users to temporally segment the video and download the resulting clips. They also have the option of downloading the videos metadata in multiple formats. Users may save interesting videos by adding them to a particular collection, or by marking them as a favourite.

Linking and related material Linking is provided in several ways to assist users in browsing the collection. Various metadata are automatically linked to queries, providing a linking by search mechanism for the video metadata. Linking by search is also provided at the similarity level, where users can click to perform quick similarity searches using search results.

Related videos and related segments are also provided. To establish links we extract from the link anchor the 10 most seldom terms with respect to the whole collection. The link targets are then determined by searching the speech representation of the segments for these terms using the lucene engine. The top returned results are returned as links.

Social features Users are able to "like," "dislike," and "favourite" videos. Statistics on how many likes/dislikes/etc. a video has are collected and made available for other users, providing an indication of how popular or controversial a video is. Statistics on less explicit actions, such as downloading a video or viewing it, are also collected and displayed. These statistics also allow for users to browse collections by the most popular videos, where popularity is determined by number of views, number of downloads, most liked, etc. Users can also add public notes to videos, and see notes that others have added.

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] Kevin McGuinness, Noel E O'Connor, Robin Aly, Franciska De Jong, Ken Chatfield, Omkar M Parkhi, Relja Arandjelovic, Andrew Zisserman, Matthijs Douze, and Cordelia Schmid, "The axes pro video search system," in *Proceedings of ICMR*, 2013, pp. 307–308.

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

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

[4] 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.

[5] 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.

[6] Hervé Jégou, Matthijs Douze, and Cordelia Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.

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