BEYOND METADATA: SEARCHING YOUR ARCHIVE BASED ON ITS AUDIO-VISUAL CONTENT

T. Tommasi[1], R. Aly[2], K. McGuinness[3], K. Chatfield[4], R. Arandjelovic[4], O. Parkhi[4], R. Ordelman[5], A. Zisserman[4], T. Tuytelaars[1]

[1]KU Leuven, ESAT-PSI, iMinds, Belgium
[2]University Twente, The Netherlands
[3]Dublin City University, Ireland
[4]University of Oxford, United Kingdom
[5]Netherlands Institute for Sound and Vision, The Netherlands

ABSTRACT

The EU FP7 project AXES aims at better understanding the needs of archive users and supporting them with systems that reach beyond the state-of-the-art. Our system allows users to instantaneously retrieve content using metadata, spoken words, or a vocabulary of reliably detected visual concepts comprising places, objects and events. Additionally, users can query for new concepts, for which models are learned on-the-fly, using training images obtained from an internet search engine. Thanks to advanced analysis and indexation methods, relevant material can be retrieved within seconds. Our system supports different types of models for object categories (e.g. “bus” or “house”), specific objects (landmarks or logos), person categories (e.g. “people with moustaches”), or specific persons (e.g. “President Obama”). Next to text queries, we support query-by-example, which retrieves content containing the same location, objects, or faces shown in provided images. Finally, our system provides alternatives to query-based retrieval by allowing users to browse archives using generated links. Here we evaluate the precision of the retrieved results based on textual queries describing visual content, with the queries extracted from user testing query logs.

INTRODUCTION

Audio-visual archives traditionally retrieve information based on manually created metadata, which is costly, subjective and coarse-grained (e.g. indexing at document level rather than shot or frame level in case of video). Especially for the casual user, searching based on such metadata may also be unintuitive since he/she may have limited knowledge of the repositories content. By directly exploiting the audio-visual information, content-based retrieval methods, on the other hand, manage to succeed even in the absence of metadata. As such, they have large potential to make retrieval more effective, they come at low cost and yield objective results at the most detailed level (e.g. keyframes). Nevertheless the existing implementations of content-based retrieval methods are somehow limited. Current systems often employ cues that have little meaning to users (e.g. colour distributions). In academic prototypes archives are usually indexed on the basis of a limited set of predefined keywords or concepts that strongly constrain users in

expressing their needs. We refer the reader to [Snoek2009] for a general overview on content-based retrieval.

The mission of the FP7 project AXES[1] is developing tools that provide various types of end users with new engaging ways to interact with audio-visual libraries, helping them to discover, browse, navigate, and search the archives in a more flexible way. This naturally promotes the use of audio-visual archives and opens their cultural wealth to the public.

We target specifically media professionals, academic researchers and journalists. Through various user studies, we found that media professionals search the archive for content on a daily basis, mostly for re-use. The main interest of academic researchers and journalists is instead in looking for new material, browsing and keeping track of their search results for future investigations. We built a tailored interface for each of these two user groups based on their requirements and our technology. Screenshots of the user interfaces built for media professionals ('AXES-PRO') and researchers ('AXES-RESEARCH') can be found in Figures 1 and 2. For more details on the different requirements of these two groups of end users, and how this affects the user interface design, we refer to [Kemman2012].

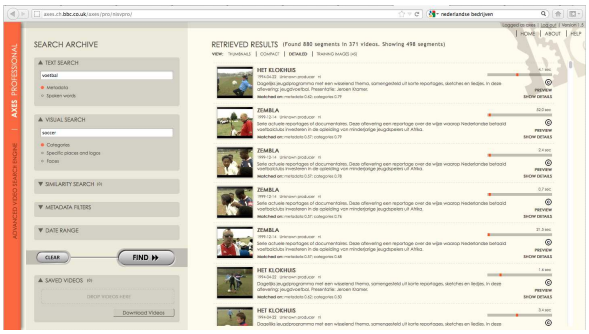

Figure 1: The AXES PRO interface

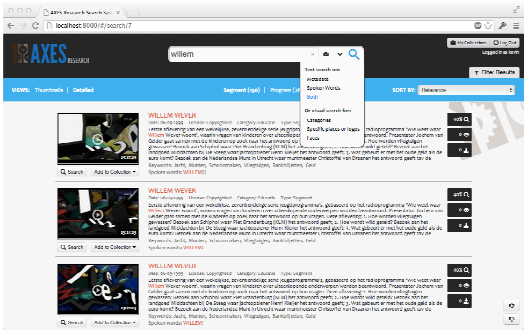

Figure 2: The AXES RESEARCH interface

For all user groups, search based on a textual query remains a dominant way of retrieving content, and a simple Google-like interface seems to be preferred by many. However, rather than just matching the search queries to the provided metadata, the AXES system, through the combination of *offline* and *online* data processing, allows direct search in the audio-visual content based on the textual query provided by the user. For example, by typing "frog" as query term, the system retrieves not only videos that have "frog" in the corresponding metadata (*metadata-search*), but also videos where "frog" is mentioned in the audio (*speech-search*) or videos that show one or more frogs (*visual search*). For the visual search, the underlying machinery is different depending on whether the query term refers to an object or scene category (e.g. "frog" or "kitchen" – referred to as *category-search*), to a person or person characteristic (e.g. "President Obama" or "people with moustaches" – referred to as *faces-search*), or to landmarks and logos (e.g. "the Eiffel Tower" or "the Coca Cola logo" – referred to as *instance-search*). For all of them, we can either use pre-trained models or learn a new model *on-the-fly* based on training images retrieved from the web using a standard image search engine. Best results are obtained when the user specifies which type of search to launch (i.e. metadata, speech, categories, faces, or instances).

[1] Full consortium includes partners from ERCIM, KU Leuven, Airbus Defense Systems, BBC, Deutsche Welle, Dublin City University, Erasmus University Rotterdam, Fraunhofer Institute, INRIA, Netherlands Institute for Sound and Vision, Technicolor, University of Oxford, and University of Twente. – see http://www.axes-project.eu

Apart from entering keywords, a search can also be based on one or more images, either a keyframe from a video found in another search, or an image uploaded by the user. The currently supported options for such similarity search are instance search (i.e. the same scene or object, e.g. a house façade or book cover) and face search. Finally, we also provide hyperlinking, i.e. linking anchors in a video fragment with target data found elsewhere in the archive.

In the rest of this paper, we present more details on the architecture of our system focusing on the text-based search. We start with the different steps in our indexing pipeline, and continue by describing the audio-visual components. Finally, we report on a small quantitative and qualitative evaluation of the search methods.

AXES SYSTEM ARCHITECTURE - SEARCHING BASED ON AUDIO-VISUAL CONTENT

Considering that an archive may grow over time, we define our system such that new content (videos or metadata) can always be added and elaborated. This includes indexing as well as a set of pre-processing steps needed to ensure fast interaction at runtime for the on-the-fly components.

The offline pipeline consists of the following steps: i) format conversion and normalization, ii) metadata indexing, iii) audio analysis, iv) video segmentation, v) pretrained object category classification, vi) pretrained event detection, vii) indexing for on-the-fly category search, viii) indexing for on-the-fly face search, ix) indexing for on-the-fly instance search, x) indexing for instance similarity search, and xi) indexing for face similarity search.

At query time the retrieval score produced by each component for every data sample (either videos or shots) is passed online to our central LIMAS service that merges the values into a single confidence score and provides the final rank list.

Audio Analysis

Most part of the audio content of the archive data is generally not reported in the metadata and thus cannot be retrieved. Speech recognition has proven to be extremely valuable in this sense. A speech-to-text tool automatically converts the spoken audio to a time-coded transcript, which can be searched, read and downloaded. We have integrated automatic speech recognition for Dutch, German and English.

Video Segmentation

Video segmentation is relevant for two reasons. First, users are often not interested in full videos, but rather in video fragments. This holds especially for media professionals who want to re-use a particular fragment with specific content. Second, most visual analysis is performed on keyframes only. This is much more efficient than analysing each and every frame of the video, or processing the video as a spatio-temporal volume. Only for face detection and event recognition, we work directly on the video by exploiting the temporal dimension (e.g. imposing temporal continuity). This yields better results, but imposes a higher computational burden.

On-the-fly search

All the archive images are processed in the offline stage for the extraction of visual features. A classification model can then be used to judge them and assign a ranking

score. The on-the-fly search learns a discriminative classifier at runtime. Once the user has specified a textual query, this is translated into a set of images through Google or Bing Image search.

In case of a *category search* visual features are extracted from the top-ranking 200 images returned by the search engine. These are used as positive samples to learn a linear SVM classifier. The negative training data is sourced during the offline stage, and is fixed for all queries. Features are computed for 1000 images downloaded from Google Image search using the publicly available API and the search term 'things' and 'photos'. The system follows closely the details given in [Chatfield2012], with the difference being that we used VLAD [Jegou2010] instead of bag-of-visual-words encoding.

The *face search* aims at retrieving video fragments based on the faces they contain. In the offline processing, faces are detected in every frame. Faces of the same person are then linked together within a shot to form face tracks. At the same time, nine facial features such as eyes, nose, mouth corners etc. are located within every detected face using a pictorial structure based method [Everingham2009]. These features provide landmarks for computing facial descriptors (feature vectors). The whole process of representing faces in the videos by tracks results in substantial reduction in data to be processed. As for the category search, positive images are downloaded from Google, while a set of negative images is taken from the Labeled Faces in the Wild dataset [Huang2007] and kept fixed for all the queries. The resulting system can be used for searching both for specific persons as well as people with facial attributes such as gender, facial hair, eyewear, etc. For more details on the method we refer to [Parkhi2012].

The *instance search* serves to quickly retrieve keyframes containing specific logos or places based on their visual appearance. Unlike the other two on-the-fly search systems, no discriminative classifier is trained. Instead, we rely on the matching of local features and geometric verification.

We start by retrieving the top 8 images downloaded from Google image search for the given query. For each of these images, a ranked results list is obtained based on the standard specific object retrieval approach of [Philbin2007], with some recent improvements. We use RootSIFT [Arandjelovic2012a] descriptors extracted from the affine-Hessian interest points of [Perdoch2009] and quantized into 1M visual words using approximate k-means. The system ranks images based on the term frequency inverse document frequency (tf-idf) score. This can be done efficiently through the use of an inverted file index. Spatial re-ranking is performed on the top 200 tf-idf results using an affine transformation model [Philbin2007]. Then the retrieved ranked lists are combined by scoring each image by the maximum of the individual scores obtained from each of the top 8 images. This is the MQ- Max method from [Arandjelovic2012b].

Pre-trained models

For popular queries, it pays off to learn better models offline. This has a double positive effect on the user experience. First, while the on-the-fly methods need to find a compromise between accuracy and processing speed, the pre-trained models can rely on larger and/or cleaner (curated) training datasets and higher-dimensional representations. As a consequence, they typically yield more accurate search results. Second, at runtime they can return results instantaneously whereas the on-the-fly methods need a couple of seconds to retrieve images from the internet, compute a new model and apply it to all images in the archive.

For the object *categories*, we use Fisher Vector encoding on top of a Gaussian Mixture Model on densely sampled SIFT features. We considered a first set of 1000 classes previously used for the ImageNet Large Scale Visual Recognition Challenge 2010 (ILSVRC 2010, [Deng2009]) and containing roughly 1000 images per category. We then enlarged this group by adding a new set of 539 classes, each containing more than 500 images and indicated as "popular" by ImageNet on the basis of the number of results returned by Google text search and of the word frequency in the British National Corpus. For each object category we trained a linear SVM using as negative samples a random set of 10K Flickr images.

We also provide a set of pre-trained classifiers for *events*. Here, we use a combination of various visual features (a dense colour descriptor, densely sampled SIFT features, and motion boundary histograms computed on dense trajectories), audio features (Mel-Frequency Cepstral Coefficients) as well as video OCR output and automatic speech recognition output. Models are trained for the different event categories provided as part of the TrecVid evaluations. For more details on the event detection, we refer to [Aly2013].

QUANTITIVE EVALUATION

We evaluated our system on different datasets. Here, we report results obtained with roughly 3000 hours of content provided by the Netherlands Institute for Sound and Vision. This dataset consists of 4522 videos, which were segmented into 1,6 Million shots, represented by almost 5M keyframes. The content is mostly Dutch broadcast material, including documentaries, talk shows, etc.

From an initial set of 10K user-testing query logs, we manually selected a subset of queries that seemed well suited for visual search (e.g. not abstract). 20 of these correspond to queries for which we also have a pre-trained model available (including, e.g., “truck”, “butterfly”). We use these to compare the results of the on-the-fly category search with those obtained using the pre-trained models. Moreover, we selected another 10 queries of object or scene categories for which no pre-trained model is available in our system (e.g. “christmas tree”, “chicken shed”, “corn field”). The wide variety of search terms we found in the query logs, confirms our intuition that it is virtually impossible to have pre-trained models for each and every category. This motivates the use of an on-the-fly scheme. We also considered 8 queries that seemed well suited for the on-the-fly instance search (e.g. “Turkish Airlines logo”), and a set of 18 queries that correspond to people (e.g. “Boris Yeltsin”, “Steve Balmer”) for the on-the-fly faces search.

For each of these queries, we retrieve the top 40 keyframes (video fragments) from the archive. We then visually evaluate whether they are indeed relevant to the search query or not. This allows us to report precision@20 and precision@40. Unfortunately, since we do not have complete ground truth for this dataset, we cannot report any data on recall. However, the precision scores allow comparing the different services to one another.

Evaluation of the on-the-fly components

First, we report results obtained with visual search (using the appropriate on-the-fly search service) considering as baseline the metadata search results – see Table 1. The metadata are usually provided at the level of a full video but only the first keyframe of the retrieved video is shown. The relatively high scores obtained by metadata search for instances are all due to only one query term, that actually corresponds to the name of a tv program and the shown keyframes are annotated as relevant because they contain the program logo.

Overall it is clear that in our simple image retrieval setup a content-based search is to be preferred. However, at the same time the outcome of our analysis also suggests that more advanced ways to investigate the metadata information and provide related results should be considered.

While the numbers may seem rather low, it is important to put them into context: with a value of 10% the on-the-fly system allows a user to find on average 2 keyframes relevant to a query out of the full set of almost 5M keyframes, by visually inspecting only 20 of them.

For the faces, we found that for some queries (mostly corresponding to Dutch politicians), results were almost perfect, while other people that probably appear less frequently in the corpus yield very low results.

	On-the-fly		Metadata	
	Precision@20	Precision@40	Precision@20	Precision@40
Categories	10,68 %	10,08 %	2,93 %	1,72 %
Instances	17,5 %	13,25 %	10,65 %	8,75 %
Faces	25,83 %	21,39 %	2,77 %	2,77 %

Table 1: Retrieval precision results: results obtained with on-the-fly and metadata search.

Comparison on-the-fly methods vs pre-trained models

Additionally, we also compare the results obtained with the on-the-fly category search with those obtained with the pre-trained models – see Table 2. Clearly, the pre-trained models give much more accurate results. This indicates that it's indeed worth the effort to collect good training images and train good models for popular queries. One of the reasons why the on-the-fly methods fail, is that the images found by Google may look quite different from those found in the archive (e.g. a single object on a white background). Our current on-the-fly methods cannot cope with such domain shift.

The histogram in Figure 3 reports the number of relevant images out of the top 20 retrieved keyframes for the category search. For some of the queries all the modalities fail and this gives a further indication of the difficulty of the task.

	On-the-fly		Pre-trained	
	Precision@20	Precision@40	Precision@20	Precision@40
Categories	12,25%	11,25%	50,75%	41,25%

Table 2: Retrieval precision results: on-the-fly vs pre-trained model comparison.

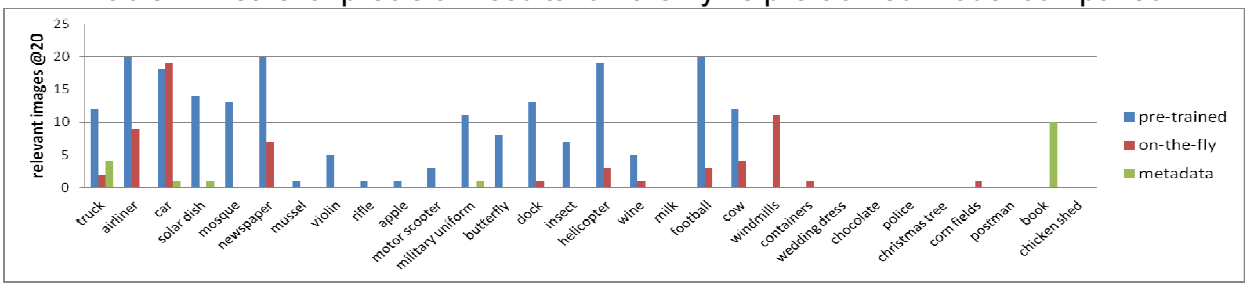

Figure 3: Category search results: for the first 20 classes pre-trained classifiers are available.

QUALITATIVE EVALUATION

We show here the visual results obtained for some of the queries used in the quantitative evaluation. In particular Figure 4-left presents the first five retrieved images for the face query “Hans Wiegel” (top line) and the instance query “Lama's” (bottom line) when using the on-the-fly methods, while Figure 4-right shows the corresponding results obtained when searching over the metadata. The on-the-fly method provides better results with respect to the metadata in the first case, but the behaviour inverts in the second case when the query is the name of a tv program (which is usually available as metadata) easily recognizable by its logo.

In Figure 5 we compare instead the result of the query “newspaper” (top line) and “helicopter” (bottom line) when using their pre-trained visual model (left) with the results obtained through the on-the-fly method (right). Here the improvement brought by a better defined pre-trained model is noticeable.

Figure 4: Query “Hans Wiegel” and “Lama's”. Left on-the-fly, right metadata.

Figure 5: Query “newspaper” and “helicopter”. Left pre-trained, right on-the-fly.

CONCLUSIONS AND FUTURE WORK

The AXES system was successfully implemented at the Netherlands Institute for Sound and Vision (in Hilversum), and the BBC (in London). Deployment at the user partners’ sites proved local installation is feasible, and showed flexibility and adaptability, as well as allowing user partners to actually try and demonstrate it in their own environment.

In addition to regular end-user testing, the individual components of our AXES system (e.g. event detection) took part in benchmark evaluations, such as in the TrecVid (video retrieval evaluation) [Aly2013], MediaEval (multimedia benchmark evaluations) [Eskevich2013], and THUMOS (large scale action recognition challenge) [Wang2013]. In all cases, the AXES team achieved excellent results.

As future work we plan to tailor our system also for home users besides the already targeted media professionals and academic researches. Moreover, given their high precision we are extending the set of pre-trained models from the existing object categories and events to faces and object instances. In addition we are finalizing a tool to automatically analyze any textual query and recognize which components are the best

suited. Visual, audio and metadata searches can then be run in parallel. This frees the user from having to make a, possibly unintuitive, choice of which search modality to use and potentially improves the final results.

REFERENCES

[Kemman2012] M. Kemman, M. Kleppe, H. Beunders, Who are the users of a Video search system? Classifying a Heterogeneous group with a profile Matrix. WIAMIS 2012.

[Deng2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. Computer Vision and Pattern Recognition, 2009.

[Aly2013] R. Aly, R. Arandjelovic, K. Chatfield, M. Douze, B. Fernando, Z. Harchaoui, K. McGuinness, N. E. O'Conner, D. Oneata, O. M. Parkhi, D. Potapov, J. Revaud, C. Schmid, J. Schwenninger, D. Scott, T. Tuytelaars, J. Verbeek, H. Wang, A. Zisserman, The AXES submissions at TrecVid 2013. TrecVid 2013.

[Everingham2009] M. Everingham, J. Sivic, and A. Zisserman, Taking the bite out of automatic naming of characters in TV video. Image and Vision Computing, 27(5), 2009.

[Huang2007] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[Parkhi2012] O. M. Parkhi, A. Vedaldi, and A. Zisserman, On-the-fly specific person retrieval. Workshop on Image Analysis for Multimedia Interactive Services, 2012.

[Philbin2007] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, Object retrieval with large vocabularies and fast spatial matching. Computer Vision and Pattern Recognition, 2007.

[Arandjelovic2012a] R. Arandjelovic and A. Zisserman, Three things everyone should know to improve object retrieval. Computer Vision and Pattern Recognition, 2012.

[Perdoch2009] M. Perdoch, O. Chum, and J. Matas, Efficient representation of local geometry for large scale object retrieval. Computer Vision and Pattern Recognition, 2009.

[Arandjelovic2012b] R. Arandjelovic and A. Zisserman, Multiple queries for large scale specific object retrieval. British Machine Vision Conference, 2012.

[Eskevich2013] M. Eskevich, R. Aly, R. Ordelman, S. Chen and G. J. Jones, The Search and Hyperlinking Task at MediaEval 2013. Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, 2013.

[Wang2013] H. Wang and C. Schmid, LEAR-INRIA submission for the THUMOS workshop, THUMOS notebook paper, 2013.

[Snoek2009] C. G. M. Snoek and M. Worring, Concept-based video retrieval. Found. Trends Inf. Retriev., vol. 4, no. 2, 2009.

[Chatfield2012] K. Chatfield, A. Zisserman, VISOR: Towards On-the-Fly Large Scale Object Category Retrieval. Asian Conference on Computer Vision, 2012.

[Jegou2010] H. Jegou, M. Douze, C. Schmid, P. Perez, Aggregating local descriptors into a compact image representation. Computer Vision and Pattern Recognition, 2010.