Open Up Cultural Heritage in Video Archives with *Mediaglobe*

Christian Hentschel, Johannes Hercher, Magnus Knuth,
Johannes Osterhoff, Bernhard Quehl, Harald Sack,
Nadine Steinmetz, Jörg Waitelonis, Haojin Yang
{givenname.lastname}@hpi.uni-potsdam.de
Hasso Plattner Institute for Software Systems Engineering
Prof.-Dr.-Helmert-Str. 2–3, 14482 Potsdam, Germany

Abstract:
Film, video, and TV have become a predominant medium, but most audiovisual (AV) material being part of our cultural heritage is kept in archives without the possibility of appropriate access for the public. Although digitalization of AV objects in conjunction with AV analysis is making progress, content-based retrieval remains difficult because of the so called semantic gap. The *Mediaglobe* project is focussed on digitalization, indexation, preservation and exploitation of historical AV archives. In this context, we show how traditional AV analysis is complemented with semantic technologies and user-generated content to enable content-based retrieval exposing contentual dependencies to promote new means of visualization and explorative navigation within AV archives.

1 Introduction

Audiovisual media such as film, video, and TV have become the predominant medium of the 21th century. From its beginning in the early 1900s until today, all kind of events all around the globe have been captured on celluloid, magnetic tape, or digital hard discs. By accessing these recordings, we are able to turn back the clock and take a peek on people and places of bygone time. But, most AV material being part of our cultural heritage is kept in archives without the possibility of appropriate access for the public audience. Even worse, celluloid is subject to an aging process and deteriorates over time. Also extensive (analog) preservation technology cannot guarantee long time storage and access. Thus, more and more AV material from the early age of film will be irrecoverably lost. Therefore, governments worldwide are funding preservation and digitalization projects, such as NESTOR[1] – the German competence network for digital preservation – to prevent the loss of cultural heritage. Everyday, the stock of AV material is growing. Recently also internet and world wide web (WWW) are used for media distribution, sometimes also providing random access and online search. But the number of online available AV material is only the tip of the iceberg, while its basis remains hidden in multitudinous analogue archives and libraries. The opening of these archives for online access requires

[1] http://www.langzeitarchivierung.de/Subsites/nestor/EN/

digitalization of various analogue media formats of different quality first. But, to access AV materials by its content, metadata for the description of its structure and content are mandatory.

In this paper, we present the *Mediaglobe* semantic video search engine and introduce its workflow including video analysis, metadata generation, semantic analysis, and video search. The primary goal of the *Mediaglobe* project[2] is to develop a generally applicable and commercially efficient infrastructure for digitalization and retrieval of AV archives with an emphasis on historical documentaries. Here, we are focussing on the retrieval process and its prerequisites, leaving out the entire process of digitalization and restoration. Most important entities to be retrieved are persons, locations, organizations, and events. We describe automated AV analysis techniques and how they are complemented by manual and collaborative annotations. Text-based metadata are endorsed by formal semantic ontologies to enable cross linking of data as well as explorative search. Instead of presenting search results in a traditional linear way, we show how semantic relationships can be utilized to enable efficient visualization and navigation.

The paper is structured as follows: Section 2 gives a short overview of related work, while Section 3 introduces the *Mediaglobe* system architecture and its analysis workflow. Visual analysis including scene cut detection, text recognition and visual concept detection in video is presented in Section 4. Section 5 focusses on the semantic analysis within the *Mediaglobe* project including an approach for named entity recognition. In Section 6 we demonstrate the semantic search facilities of the *Mediaglobe* search engine and Section 7 provides an overview of the newly developed semantic search interface. Section 8 concludes the paper with a short outlook on ongoing and future work.

2 Related Work

Current web-based video platforms such as YouTube allow their users to upload content and to assign basic metadata (title, keywords, description). Although, YouTube enables the annotation of temporal video segments using fragment identifiers up to now there is no support for semantic annotation as in *Mediaglobe*. In [SH10] Steiner and Hausenblas exemplify the processing of YouTube's metadata with natural language processing tools (NLP) in order to tag existing video metadata with linked data entities, but this approach is limited to the video as a whole, whereas in *Mediaglobe* semantic annotation is time based and automatically generated via sophisticated video analysis technologies. With regard to time based semantic video annotation for enabling semantic and exploratory video search we refer to previously published in depth results [WKW$^+$10, LS11, WS11]. In [KSR09] Kobilarov et al. present the BBC's approach to join up all of its resources using linked data principles and a tailored ontology for BBC's program data. Later on, the NoTube project[3] harnessed BBC's program information by analyzing it with a NLP-tool to extract named entities and to map them to linked data resources. Again, both approaches are limited to

[2] The *Mediaglobe* project is part of the THESEUS research program funded by the German Federal Ministry for Economics and Technology, http://www.projekt-mediaglobe.de/

[3] http://notube.tv/

program data, and also consider the annotated asset as a whole rather than its temporal segments. In contrast to the BBC program data and the NoTube project in *Mediaglobe* there is no access to rich editorial materials [vAARB09]. Consequently, metadata in *Mediaglobe* has to be generated by automated video analysis techniques such as scene- and overlay-text recognition or automated speech recognition.

Mature projects including Prestospace, Europeana, and THESEUS/CONTENTUS build foundations for semantic content analysis within the cultural heritage domain. The aim of Prestospace[4] was to develop preservation technology for broadcasters. In [MBS08] Messina et al. describe tools for content-based analysis, e. g., scene-cut detection, speech to text transcription and keyframe-extraction. Moreover, named entities and categories from the BBC's program web-pages were also extracted and mapped to the proton upper-ontology [DTCP05]. *Mediaglobe* complements Prestospace's efforts, as high-level abstraction layer analysis technologies, i. e. visual concept detection was not in the scope of Prestospace. In regards of semantic analysis, video assets in *Mediaglobe* rarely come with rich accompaniment metadata as it is the case of tv programs. Therefore, background knowledge has to be aggregated that is maintained independently of the original source, i. e., the film archive. In contrast to [DTCP05, MBD06] we access the DBpedia as multilingual knowledge base to allow crosslingual search and annotation with entities that are steadily curated by the Wikipedia community. The Europeana[5] aggregates video data including supporting documents from various film archives in Europe. Prominent examples of video platforms that supply the Europeana search interface include the European Film Gateway (EFG), the videoactive platform, and EU-screen. However, these projects concentrate on content alignment and crosslingual search by harmonizing existing but heterogeneous metadata originating from broadcasters, film archives, and cinematheques [SET$^+$12]. To our best knowledge, they do not apply sophisticated content-analysis technology to leverage metadata nor perform additional time-based semantic video analysis. Within the CONTENTUS usecase of the German THESEUS research program technologies supporting the media library of the future have been developed [NLFH12]. CONTENTUS targets multimedia in all its diversity whereas *Mediaglobe* is specialized on video only. *Mediaglobe* complements the semantic video analysis technology of CONTENTUS by applying semantic graph analysis for entity mapping as well as by applying high-level abstraction layer analysis technologies, such as visual concept detection.

3 System Architecture

To process large amounts of video content, many requirements have to be met including high scalability, customizable workflow design, reusability, versioning, and compatibility with state-of-the-art standards. Since traditional video processing workflow systems focus on transcoding, delivery, publishing, and metadata management, they still lack sophisticated and semantic search to open up its content. Therefore, *Mediaglobe* is designed not only to perform standalone, but also to complement existing workflow systems by bringing

[4] http://prestospace.org/
[5] http://europeana.eu/

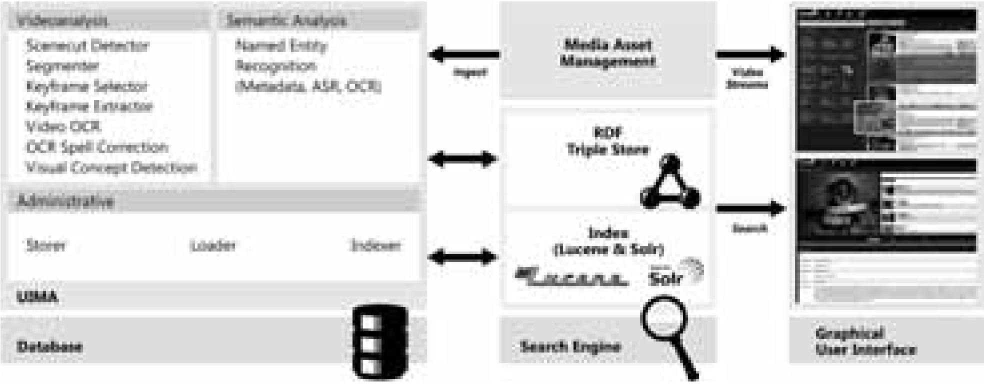

Figure 1: The overall architecture of the *Mediaglobe* system

in flexible video content analysis and semantic search features.

The *Mediaglobe* system consists of three main parts: a *Media Asset Management System* (MAM), the *Analysis Framework* and the *Video Search Engine*. The MAM maintains the video ingest, transcoding, play out streaming, rights management, and commercial exploitation. The analysis framework builds upon the Apache Unstructured Information Management Architecture (UIMA) that enables the compilation of various Analysis Engines (AE) and orchestrates the data flow between them. Work intensive components can be deployed plurally for parallel and/or distributed execution. The video search engine provides a semantic search index and serves the Graphical User Interface (GUI) to efficiently browse and search within the video data. Fig. 1 gives an overview on the main components and the overall architecture. After ingesting a new video from the MAM the workflow engine starts with scene cut detector and keyframe extractor as preprocessing for subsequent video Optical Character Recognition (OCR) and visual concept detection. The auditive tracks are processed by third-party Automatic Speech Recognition (ASR). All extracted textual information including metadata provided by the users is processed by the semantic analysis. This employs the temporal context dependent Named Entity Recognition (NER). Finally, the semantically enriched data is made available in an RDF triple store to query via SPARQL and a search engine for end user access. All essential components are introduced in the following sections, starting with the temporal decomposition of video with scene cut detection.

4 Visual Analysis

The low-level analysis of the visual content of video data provides additional source for metadata. This section describes visual analysis techniques employed in *Mediaglobe*.

4.1 Scene Cut Detection and Keyframe Selection

Structural video analysis by scene cut detection is the first step of video content analysis, indexing, and classification. Video shots are defined as a sequence of consecutive frames taken by a single camera act. Shot boundary detection provides useful information about the temporal structure of a video, which is a prerequisite to the extraction of representative keyframes or frame candidates for video OCR, visual concept detection, and to define temporal contexts for NER. We distinguish two different types of shot boundary transitions: abrupt scene changes (*hard cuts*) and gradual transitions that extend to more than a single frame (*soft cuts*) [AAM10].

Most common gradual transitions are *fade-ins* and *fade-outs*, *wipes* and *dissolves*. As the video data under consideration in the *Mediaglobe* archive restrains to hard cuts and fades, we do not consider wipe and dissolve detection here. Our approach for hard cut detection is based on a statistical method expanded on the idea of pixel differences [ALBK09]. We compute the the $L2$-norm between every 5 consecutive frames and consider a frame a hard-cut candidate if the gradient computed on the first derivative of this metric exceeds an empirically determined (adaptive) threshold. Fade-ins and fade-outs exhibit a slow decrease or increase change in illumination usually ending or starting with a black frame. Detection is based on computing the image entropy – a monotonously increasing entropy value indicates a fade-in, beginning and end frame for this transition are defined by the local minimum and maximum respectively.

Next to shot-boundary detection scene cut detection provides input data that is qualified for further content-based visual analysis of a video. In particular, the selection of meaningful keyframes per segment provides efficient means to reduce the temporal redundancy within the visual data of a video and thus the computational effort required for subsequent analysis steps, e. g., video OCR and visual concept detection. While the video OCR runs on up to 5 frames selected equidistantly per shot, the visual concept detection assumes that a single frame provides enough information about the visual content of an entire video segment, and therefore has to be chosen in a way to contain the most meaningful content.

4.2 Video OCR

Text embedded in video data is a valuable source for indexing and searching in video content. In order to retrieve textual data from video, standard OCR approaches, which focus on high resolution scans of printed documents, need to be extended to meet the requirements for OCR in video data. Our approach consists of two steps: keyframes that contain textual data are identified and subsequently, the location of text within these frames is determined in order to separate text pixels from background and to deliver this preprocessed data to a standard OCR engine. In our approach, we classify text candidates from the video data stream by applying a fast edge-based multi-scale detector and subsequently a projection-profiling algorithm on each keyframe. Frame sections containing no text are rejected by identifying regions that exhibit low edge density. Further refinement is achieved by a Stroke Width Transform (SWT) based verification procedure on each detected text candidate section. We adapted the original SWT algorithm [EOW10] in order to improve the performance [YQS12b].

Typically, video text is embedded in very heterogeneous background with low contrast, which makes it difficult to be recognized by standard OCR engines. Hence, text pixels need to be separated from background by applying appropriate binarization techniques. We have developed a novel skeleton-based approach for video text binarization [YQS12a], which consists of three steps: First, we determine the text gradient direction for each text line object by analyzing the content distribution of their skeleton maps [CPW93]. We then calculate the threshold value for seed-selection by using the skeleton map which has been created with the correct gradient direction. Subsequently, a seed-region growing procedure starts from each seed pixel and extends the seed-region in its north, south, east, and west orientations. The region grows iteratively until it reaches the character boundary.

Text binarization is concluded by converting text region data to black text pixels on white background, which is a prerequisite for the efficient application of a standard OCR software (*Tesseract*[6]) for text recognition. Moreover, an adapted version of the open source *Hunspell*[7] spell checker is applied to improve the quality of the achieved OCR results. While traditional spell checking software presumes typing errors based on a specific keyboard layout, an adapted OCR spell correction rather cares for visual similarity of the characters.

4.3 Visual Concept Detection

Next to textual data, the depicted visual concepts within a video scene provide additional information for search and classification. We employ methods for content-based image classification to automatically classify video scenes into predefined visual concept classes. Our approach follows the well-known *bag of keypoints* [CDF$^+$04] model, where local image patterns are used to describe the low-level visual features of a keyframe. In our approach, we extract Scale Invariant Feature Transform (SIFT) [Low04] features at a fixed grid on each channel of a keyframe in RGB color space. These features are used to compute a visual codebook by running a k-means clustering that provides us with a set of representative codewords. By assigning each extracted RGB-SIFT feature to its most similar codeword (or cluster center) using a most simple nearest neighbor classifier we compute a histogram of codeword frequencies that is used as keyframe descriptor. The problem of visual concept detection is approached by means of machine learning techniques. Kernel-based Support Vector Machines (SVM) have been widely used in image classification scenarios (cf. [CDF$^+$04, SW09, ZMLS06]). In order to be able to predict the category of an unlabeled keyframe the SVM classifier needs to be trained using labeled data.We consider the classification task a one-against-all approach – one SVM per given visual concept is trained to separate the keyframes from this concept from all other given concepts. Hence, the classifier is trained to solve a binary classification problem, i. e.,

[6] http://code.google.com/p/tesseract-ocr/

[7] http://hunspell.sourceforge.net/

whether or not a keyframe depicts a specific visual concept. We use a Gaussian kernel based on the χ^2 distance measure, which has proven to provide good results for histogram comparison. Following Zhang et al. [ZMLS06] we approximate the kernel parameter γ by the average distance between all training image histograms. Therefore, the only parameter we optimize in a cross-validation is the cost parameter C of the support vector classification. New keyframes can be classified using the aforementioned bag-of-words feature vectors and the trained SVM model.

The advantage of this approach is its simplicity, and its various invariances. The combination of SIFT for local image description and the bag-of-words model makes it invariant to transformations, changes in lighting and rotation, occlusion, and intra-class variations [CDF$^+$04]. By simply counting object features present in an image, missing or occluded object parts do not affect the classification accuracy in the same way as it would be the case for model-driven classification. Moreover, the approach is very generic. It can be extended to additional concepts simply by training another SVM classifier. Since the features are not adapted to the classification task they can be reused.

5 Semantic Analysis

In order to enable an explorative search within video data the textual metadata derived from visual analysis and the authorative metadata (provided by the archive) are mapped to semantic entities. For *Mediaglobe* authoritative metadata comes as persons, places, organizations, other keywords, and freetext, e. g., title, description text, while visual analysis technologies provide time-based OCR and ASR text.

For semantic annotation in *Mediaglobe* DBpedia[8] entities serve as reference mapping entities for NER. Initially, for every entity all DBpedia labels including redirects and disambiguation links are collected. Next, for every label a distance measure to the original (main) label of the entity is calculated to determine a relevance ranking for entity candidates. For instance, the label 'DEBER' of the entity 'Berlin'[9] receives a lower distance score than the labels 'Berolina' or 'City Berlin', where 'Berlin' is the original (main) label. In case a term to be mapped can be assigned to several DBpedia entities, the term is ambiguous and further context data is needed for disambiguation According to their provenance metadata are of different reliability. Therefore, we have determined a context ranking according to metadata provenance. First, the authorative metadata items are disambiguated in the order [persons, places, organizations], [keywords], [freetext], whereas already disambiguated metadata terms serve as additional context for the currently processed metadata text item. Subsequently, the time-based metadata text items are disambiguated within the context boundaries a video segment. Every entity candidate of a metadata term receives a score according to its relevance within the defined context. In the next step, cooccurrence of context text terms and the entity candidates of the currently processed metadata text term is identified with the help of the Wikipedia articles of the

[8] http://dbpedia.org/About
[9] http://dbpedia.org/resource/Berlin

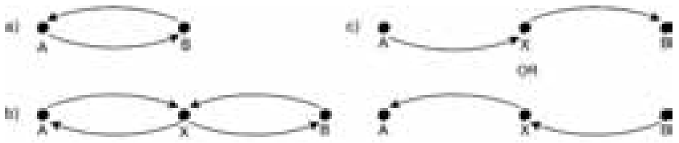

Figure 2: Three different types of Wikilinks: a) direct links, b) symmetric links through same node, c) links through a node, but not symmetric.

referenced entity candidates. The entity candidate with the highest involvement (score) within the context will be chosen as mapped entity [LS11].

In addition to cooccurrence analysis link graph analysis based on the Wikipedia page link graph is applied to identify connected components within a context. Three different types of links are considered (cf. Fig. 2). Similarly to the cooccurence analysis score the link graph score for every entity candidate is calculated according to the total count of links, but also to the count of links to different context terms taking into account how many different nodes are between a pair of entity candidates of different terms.

One of *Mediaglobe*'s key features is the collaborative and time-based annotation of video segments. To this end a customized video annotation tool has been developed, where the user is able to watch the video, stop it anytime and tag it with semantic entities or text keywords. An autosuggestion service provides the user with relevant semantic entities chosen from DBpedia for a typed text term. Once an entity has been selected the user can position the semantic tag anywhere within the video frame to achieve spatio-temporal annotation.

6 Semantic Search

Every search process starts with formulating a search query that will best express the user's information need. The type of query dictates the extent of expressiveness and how the search engine has to parse and interpret the query. Traditional search engines represent the query as independent *keywords* of interest to the user. More advanced systems allow to query with *natural language* or *formal query languages*. The first is the most expressive but also most difficult to parse and interpret by machines, the latter is hard to formulate by non-expert users but well processable by machines.

Mediaglobe combines the best of both worlds in a *hybrid* approach. It enables to search for distinct semantic entities instead of keywords and eliminates the ambiguities caused by polysemy and synonymity of natural language. Disambiguation is enforced while entering the search query via type-ahead suggestions. When the user starts to type, the indexed semantic entities matching the entered text best are all displayed. The user has to choose the desired entity from the presented candidate list. Complementing also the traditional keyword-based search, searching for distinct semantic entities improves precision and recall as well as the user experience. To enable efficient entity based search the generated semantic information has to be indexed appropriately. This implies not to store keywords

Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Term Text	john	f	kennedy johnfkennedy dbp:John_F._Kennedy	said	I	love	berlin dbp:Berlin
Type	word	word	word word URI	word	word	word	word URI
Start- / End-Offset	0 / 38	0 / 38	0 / 38 0 / 38 0 / 38	39 / 43	46 / 47	48 /52	53 / 75 53 / 75

Table 1: Search index terms with URIs and offsets after tokenizing and filtering the annotated text

and normalized terms (e. g. via stemming) only, but also to store the URIs representing the semantic entity bound to the text position the entity is determined at. Therefore, we have implemented an extension of the Lucene search engine, which includes support for annotation aware string tokenizing, word-delimiter filters, payload token filters, and efficient snippet highlighting. Before the process of indexing starts, text data is transformed into semantically annotated text containing the URIs of semantic entities being found via NER. Therefore, a simple markup '*(Label){ URI }*' is used, as e. g.:

Text to annotate: *John F. Kennedy said: 'I love Berlin!'*
Text with markup and URIs: *(John F. Kennedy){dbp:John_F._Kennedy} said: 'I love (Berlin){dbp:Berlin}!'*

Table 1 shows the final index terms after the annotated text has been processed by annotation aware tokenizer, word-delimiter filter, Lucene standard filter, lowercase filter and finally the payload processing filter, to store the URIs in payload attributes along with the index term. Searching for the URI 'dbp:John_F._Kennedy' would result in returning all documents containing only the specified entity. Traditional full text search is supported as already existing Lucene feature. The ranking of documents is based on the standard TF/IDF-based Lucene scoring.

7 The Semantic Search Interface of *Mediaglobe*

While most keyword-based search systems are optimized to narrow down huge data spaces to the most suitable results and present them in 'ten blue links', semantic exploratory search in *Mediaglobe* additionally aims at finding results, which are not considered to be related at first glance. Semantic exploratory search is based on facet entities and content-based recommendations, enabling the user to better refine and broaden search queries [WKW$^+$10]. Thus, the result space has to distinct from the classical result list of keyword based information retrieval systems. Our interface design objective was to support its users with a quick feedback on selected facet entities that encourages the exploration of the provided results. Considering Fig. 3, the layout of the *Mediaglobe* user interface is arranged in a *Search Area* and a *Result Area*. The *Search Area* contains an input field, and a list of confirmed semantic entities to its right. These entities determine the current

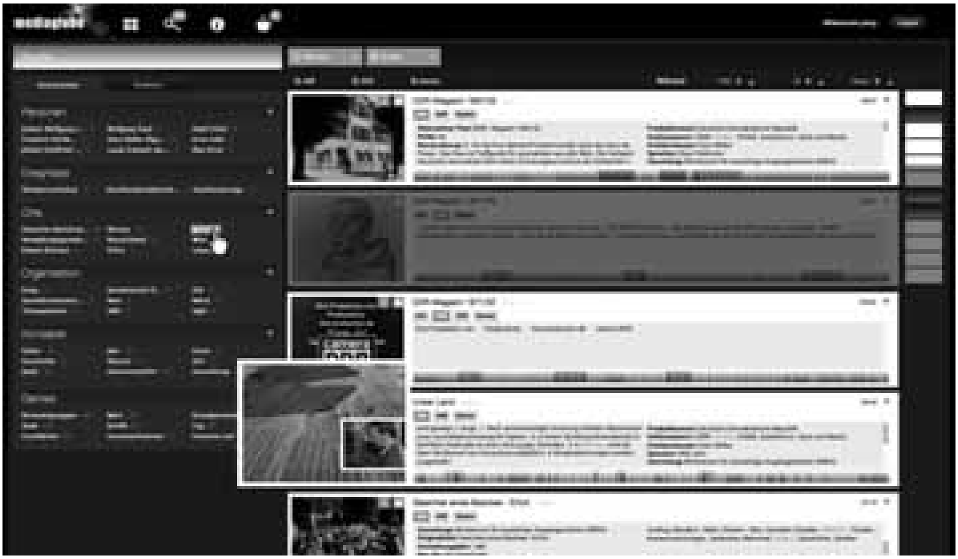

Figure 3: The user interface of *Mediaglobe*: When *brushing* over a facet on the left, the *linked* result tiles and their pagination are updated instantly

search result in the *Search Area* below, where the search results are represented via content snippets. A vertical pagination on the right represents all videos of the current result space. The facet list on the left of the search result contains the facet filters in categories for Persons, Events, Places, Organizations, Concepts, and visual Genres.

The result tiles in the middle of the layout contain the individual videos; they indicate the title of the video, a video thumbnail, and the *Snippet Container*. In case, results derived from ASR, OCR, and visual concept detection, the respective container is captioned with tabs, each of which switching to the corresponding sections and containing extracted ASR or OCR results or a list of detected visual concepts. A timeline below this container exposes the automatically generated temporal segmentation of the videos with highlighted segments that fit the currently selected entities. When users hover over the timeline, the thumbnail of the result tile is updated with an image from the respective segment. This quick video preview provides feedback regarding the structure of a video and leads to simple and concise content-based comparability of different videos. In order to achieve an interaction principle known as ‘Brushing and Linking’ [OWJS11] the search facets on the left are *linked* to the search results in the main area of the layout and the pagination on its right: when the user *brushes* over one of the facet filters, the video result tiles and their equivalents in the pagination that do not match the criteria are grayed out. By making the facets available for brushing, and linking them to the representative result tiles with the pagination, the proposed interface paradigm provides valuable feedback on how the result set will be changed according to the next distinct selection by the user.

8 Conclusion & Future Work

In this paper, we have presented the *Mediaglobe* semantic video search engine, which provides efficient access to AV archives focussing on historical documentaries. We have shown how we combine methods for visual analysis into a single workflow for automatic metadata generation. We have furthermore presented sophisticated approaches for semantic analysis supported by formal semantic ontologies to enable cross linking of data and explorative search. Finally, we have introduced a novel search interface, which joins the developed technologies and enables efficient visualization and navigation to support the user in his task of retrieving relevant data as well as exploring the archive. Due to copyright regulations on the underlying AV data it is unfortunately not possible to grant public access to the *Mediaglobe* demonstrator. However, a screencast showing the core features and presented technologies is available online[10].

Future work will improve the search ranking by incorporating semantic relations between entities. This enables to rank documents based on *semantic similarity* and not on syntactic similarity as done by traditional search engines. Our future work will also focus on enhanced semantic analysis techniques comprising more ontological information from different sources. This will increase the precision of entity assignments and improve the semantic relatedness between archive items. Furthermore, we plan to interconnect additional elements of the interface via 'brushing and linking'.

References

[AAM10] A. M. Amel, A. B. Abdelali, and A. Mtibaa. Video shot boundary detection using motion activity descriptor. *Jrnl. of Telecommunications*, 2(1):54–59, 2010.

[ALBK09] D. Adjeroh, M. C. Lee, N. Banda, and U. Kandaswamy. Adaptive edge-oriented shot boundary detection. *Jrnl. Image Video Process.*, pages 5:1–5:13, January 2009.

[CDF$^+$04] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, and D. Maupertuis. Visual Categorization with Bags of Keypoints. In *WS. on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[CPW93] C.H. Chen, L.F. Pau, and P.S.P. Wang. *Handbook of Pattern Recognition and Computer Vision*. World Scientific, 1993.

[DTCP05] M. Dowman, V. Tablan, H. Cunningham, and B. Popov. Web-assisted annotation, semantic indexing and search of television and radio news. In A. Ellis and T. Hagino, editors, *WWW*, pages 225–234. ACM, 2005.

[EOW10] B. Epshtein, E. Ofek, and Y. Wexler. Detecting Text in Natural Scene with Stroke Width Transform. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2963–2970, 2010.

[KSR09] G. Kobilarov, T. Scott, and Y. Raimond. Media Meets Semantic Web — How the BBC Uses DBpedia and Linked Data to Make Connections. In *Proc. of the 6th European*

[10] http://bit.ly/mediaglobe

Semantic Web Conf. on The Semantic Web: Research and Applications, ESWC 2009 Heraklion, pages 723–737, Berlin, Heidelberg, 2009. Springer-Verlag.

[Low04] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. Jrnl. of Computer Vision*, 60(2):91–110, November 2004.

[LS11] N. Ludwig and H. Sack. Named Entity Recognition for User-Generated Tags. In *Proc. of the 8th Int. WS. on Text-based Information Retrieval*. IEEE CS Press, 2011.

[MBD06] A. Messina, L. Boch, and G. Dimino. Creating Rich Metadata in the TV Broadcast Archives Environment: The PrestoSpace Project. In *Int. Conf. on Automated Production of Cross Media Content for Multi-Channel Distribution*, pages 193–200, Los Alamitos, CA, USA, 2006. IEEE Comp. Soc.

[MBS08] A. Messina, W. Bailer, and P. Schallauer. Content Analysis Tools – Deliverable 15.4. Technical report, RAI, 2008.

[NLFH12] J. Nandzik, B. Litz, and N. Flores-Herr. CONTENTUS—technologies for next generation multimedia libraries. *Multimedia Tools and Applications*, pages 1–43, 2012. 10.1007/s11042-011-0971-2.

[OWJS11] J. Osterhoff, J. Waitelonis, J. Jäger, and H. Sack. Sneak Preview? Instantly Know What To Expect In Faceted Browsing. 2011.

[SET$^+$12] N. Simou, J.-P. Evain, V. Tzouvaras, M. Rendina, N. Drosopoulos, and J. Oomen. Linking Europe's Television Heritage. In J. Trant and D. Bearman, editors, *Proc. Museums and the Web 2011*. Archives & Museum Informatics, Toronto, 2012.

[SH10] T. Steiner and M. Hausenblas. SemWebVid – Making Video a First Class Semantic Web Citizen and a First Class Web Bourgeois. In *9th Int. Semantic Web Conference (ISWC 2010)*, 2010.

[SW09] C.G.M. Snoek and M. Worring. Concept-Based Video Retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2009.

[vAARB09] C. van Aart, L. Aroyo, Y. Raimond, and D. Brickley. The NoTube Beancounter: Aggregating User Data for Television Programme Recommendation. In *Social Data on the Web (SDoW2009)*. CEUR WS. Proc. Vol. 520, 2009.

[WKW$^+$10] J. Waitelonis, M. Knuth, L. Wolf, J. Hercher, and H. Sack. The Path is the Destination – Enabling a New Search Paradigm with Linked Data. In *Proc. Linked Data in the Future Internet at the Future Internet Assembly, Ghent, Belgium, CEUR WS. Proc.*, volume 700, dec 2010.

[WS11] J. Waitelonis and H. Sack. Towards exploratory video search using linked data. *Multimedia Tools and Applications*, pages 1–28, 2011. 10.1007/s11042-011-0733-1.

[YQS12a] H-J. Yang, B. Quehl, and H. Sack. A skeleton based binarization approach for video text recognition. In *Proc. IEEE Int. WS. on Image Analysis for Multimedia Interactive Service (to appear)*, Dublin, Ireland, May 2012. IEEE Comp. Soc.

[YQS12b] H-J. Yang, B. Quehl, and H. Sack. Text detection in video images using adaptive edge detection and stroke width verification. In *Proc. of 19th Int. Conf. on Systems, Signals and Image Processing (IWSSIP)*, Vienna, Austria, April 11–13 2012. IEEE Comp. Soc.

[ZMLS06] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *Int. Jrnl. of Computer Vision*, 73(2):213–238, September 2006.