International Journal on Digital Libraries (2019) 20:167–183
https://doi.org/10.1007/s00799-018-0236-z

Content-based video retrieval in historical collections of the German Broadcasting Archive

Markus Mühling[1] · Manja Meister[3] · Nikolaus Korfhage[1] · Jörg Wehling[3] · Angelika Hörth[3] · Ralph Ewerth[2,4] · Bernd Freisleben[1]

Received: 31 January 2017 / Revised: 10 February 2018 / Accepted: 19 February 2018 / Published online: 8 March 2018

Abstract
The German Broadcasting Archive maintains the cultural heritage of radio and television broadcasts of the former German Democratic Republic (GDR). The uniqueness and importance of the video material fosters a large scientific interest in the video content. In this paper, we present a system for automatic video content analysis and retrieval to facilitate search in historical collections of GDR television recordings. It relies on a distributed, service-oriented architecture and includes video analysis algorithms for shot boundary detection, concept classification, person recognition, text recognition and similarity search. The combination of different search modalities allows users to obtain answers for a wide range of queries, leading to satisfactory results in short time. The performance of the system is evaluated using 2500 h of GDR television recordings.

Keywords German Broadcasting Archive · Automatic content-based video analysis · Content-based video retrieval · Deep learning

1 Introduction

In recent years, deep learning methods, in particular deep convolutional neural networks, have led to breakthroughs in many computer vision fields. Integrating deep learning methods as novel video retrieval technologies into media archives offers new potentials in accessing, searching and browsing the data stored in digital video libraries [2,6,27]. In particular, content-based analysis and retrieval in large collections of scientific videos is an interesting field of research. Examples are Yovisto[1], ScienceCinema[2] and the TIB|AV portal[3] of the German National Library of Science and Technology (TIB). The latter provides access to scientific videos based on speech recognition, visual concept classification and video OCR (optical character recognition) [19,34]. The videos of this portal stem from the fields of architecture, chemistry, computer science, mathematics, physics, and technology/engineering.

The German Broadcasting Archive (DRA) in Potsdam–Babelsberg provides access to another valuable collection of scientifically relevant videos. It encompasses significant

✉ Markus Mühling
muehling@informatik.uni-marburg.de

Manja Meister
manja.meister@dra.de

Nikolaus Korfhage
korfhage@informatik.uni-marburg.de

Jörg Wehling
joerg.wehling@dra.de

Angelika Hörth
angelika.hoerth@dra.de

Ralph Ewerth
ralph.ewerth@tib.eu

Bernd Freisleben
freisleb@informatik.uni-marburg.de

1 Department of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str. 6, 35032 Marburg, Germany

2 German National Library of Science and Technology (TIB), Welfengarten 1B, 30167 Hannover, Germany

3 German Broadcasting Archive, Marlene-Dietrich-Allee 20, 14482 Potsdam, Germany

4 L3S Research Center, Leibniz Universität Hannover, Appelstraße 4, 30167 Hannover, Germany

[1] http://www.yovisto.com.

[2] http://www.osti.gov/sciencecinema.

[3] http://av.tib.eu.

parts of the audio-visual tradition in Germany and reflects the development of German broadcasting before 1945 as well as radio and television of the former German Democratic Republic (GDR). The DRA was founded in 1952 as a charitable foundation and joint institution of the Association of Public Broadcasting Corporations in the Federal Republic of Germany (ARD). In 1994, the former GDR's radio and broadcasting archive was established. The archive contains film documents of former GDR television productions from the first broadcast in 1952 until its cessation in 1991. It includes a total of around 100,000 broadcasts, such as: contributions and recordings of the daily news program *Aktuelle Kamera*; political magazines such as *Prisma* or *Der schwarze Kanal*; broadcaster's own TV productions including numerous films, film adaptations and TV series productions such as *Polizeiruf 110*; entertainment programs (e.g., *Ein Kessel Buntes*); children's and youth programs (fairy tales, *Elf 99*); as well as advice and sports programs. Access to the archive is granted to scientific, educational and cultural institutions, to public service broadcasting companies and, to a limited extent, to commercial organizations and private persons. The video footage is often used in film and multimedia productions. Furthermore, there is a considerable international research interest in GDR and German-German history. Due to the uniqueness and importance of the video collection, the DRA is the starting point for many scientific studies. International scientists, particularly from the USA and UK, followed by the Netherlands, Japan, Sweden and Switzerland, use the DRA for their research in the fields of psychology, media, social, political or cultural science. These studies are, for example: *Heavies in East Germany* (Humboldt University Berlin), *Space Travel in the GDR* (Harvard University, USA), *The Jewish in TV* (Ludwig Maximilian University of Munich), *Socialism on the Screen* (Loughborough University, UK), *Self-made in Consumer Society* (University of Mannheim), and *Child and Youth Education in Fictional Subjects* (Shizuoka University, Japan). The DRA is answering a wide range of research requests concerning the life of GDR citizens and social perceptions. The number of comprehensive and time-consuming requests is considerably increasing, e.g., *youth fashion in the GDR, especially for punks and bluesers*; *living in East Germany, in particular home furnishings from Deutsche Werkstätten Hellerau*; or *the socialist city as a model of urban development in the GDR, specifically pictures including socialist classicism, buildings made with precast concrete slabs, demolition and spectacular buildings*.

Due to the time-consuming task of labeling videos manually, human annotations focus on larger video sequences and contexts. Furthermore, finding similar images in large multimedia archives is manually infeasible. Thus, the DRA aims to digitize and index the entire video collection to facilitate search in videos.

In this paper, we present a novel content-based video retrieval system for searching in historical collections of GDR television recordings. It contains algorithms for visual concept classification, similarity search, person recognition, and text recognition. Apart from applying content-based video retrieval to a unique and very important historical video collection, the paper makes the following contributions:

- A novel GDR specific lexicon of 91 concepts including, for example, *Trabant*, *GDR emblem*, *military parade*, *optical industry*, or *community policeman*, is used for automatic annotation.
- A deep convolutional neural network (CNN) is extended to perform multi-label concept classification; the results are compared to a Bag-of-Visual-Words (BoVW) approach.
- A novel, fast similarity search approach for large-scale video retrieval is presented.
- The performance of the system is evaluated using 2500 h of GDR television recordings.

The content-based video retrieval system is used to complement human annotations and to support users in finding relevant video shots. In contrast to manual annotations, content-based video analysis algorithms provide a more fine-grained analysis, typically based on video shots.

The paper is organized as follows: Section 2 describes the video retrieval system, including the digitization process, the content-based video analysis algorithms, the service-oriented architecture and the workflows. In Sect. 3, the results on the GDR television recordings are presented as well as a comparison of a Bag-of-Visual-Words approach and deep multi-label CNNs in the field of concept classification. Section 4 concludes the paper and outlines areas for future research.

2 A content-based video retrieval system

In this section, a content-based video retrieval system to support search in historical GDR television recordings is presented. Its aim is to automatically assign semantic tags to video shots for the purpose of facilitating content-based search and navigation.

Figure 1 shows an overview of the developed video retrieval system. First, the videos are digitized and preprocessed. The preprocessing step mainly consists of shot boundary detection with additional tasks, such as video transcoding or thumb generation, which are required for later visualization purposes. Based on video segmentation, the following automatic content-based video analysis algorithms are applied: concept classification, similarity search, person and text recognition. The resulting metadata are written to a database. Given this semantic index, arbitrary search queries

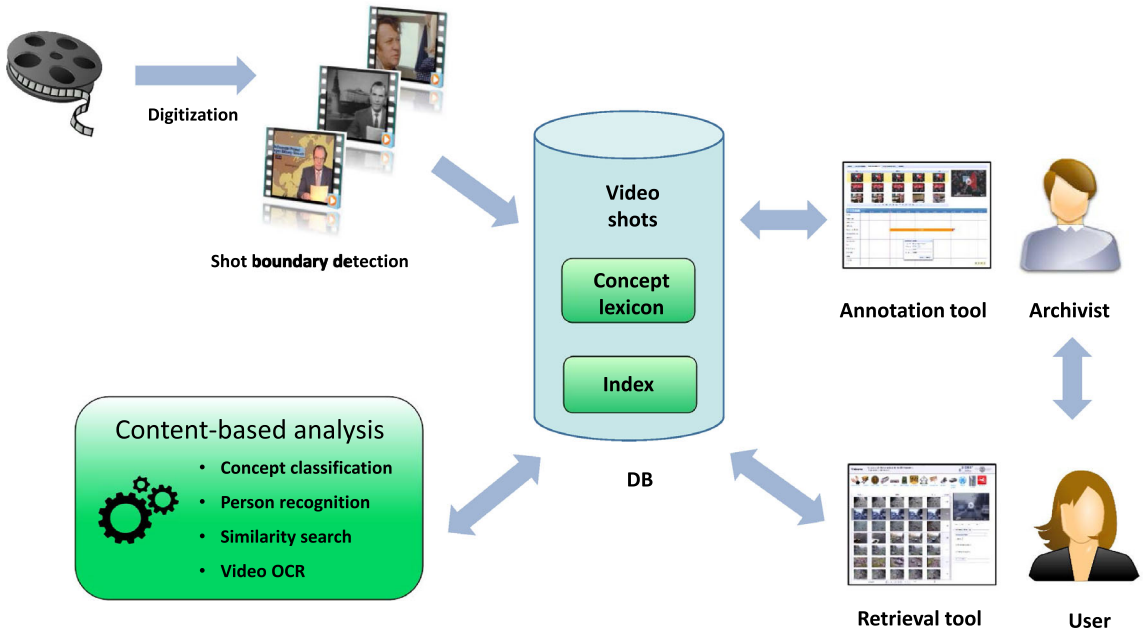

Fig. 1 Video retrieval system

can be processed efficiently. The query results are returned to the user as a list of video shots, ranked according to the probability of the presence of the desired content. Due to the large amount of video data and the associated computational requirements, distributed content-based video analysis is performed in a service-oriented architecture.

In the following, the digitization process, the analysis algorithms, the service-oriented architecture and the workflows are described.

2.1 Video digitization

Overall, 3100 h of the analog, historical GDR television recordings were digitized. The digitized material was selected from the large amount of available analog video material according to its relevance for research. In particular, it consists of socio-political magazines, the daily news program *Aktuelle Kamera* and several TV productions. The collection provides a wide range of themes and reflects everyday life in the former GDR.

The digitization process was carried out both by the DRA itself and by an external provider. In a manual preprocessing step, the tapes were technically prepared and the content was reconciled using the FESAD database of the DRA. FESAD ("Fernseharchivdatenbank") is a TV database jointly used in the ARD, which is a consortium of public-law broadcasting institutions of Germany. For the external digitization, technical parameters such as the video format (e.g., Betacam SP), the duration, and the time codes were reviewed. About 100 h of video data were digitized by the DRA. Depending on the video format, different digitization devices were used, such as an AVID workstation, DVS Fuze 5.10, DVS Venice 3.2 or Digital Vision Phoenix 2015. Finally, the digitized videos were corrected with respect to their colors.

However, most of the video tapes were digitized by the worldwide unique automated digital archiving system ADAM (Automated Digital Archive Migration). ADAM was developed by the Swiss JORDI AG[4] in collaboration with the WDR media group[5] according to archival requirements. The core of the system is an industrial robot that transfers the content of the video tapes automatically into a digital file-based archive. It grabs video tapes from a carousel and places them in a pool of up to 740 parking slots. The tapes are passed to a video tape recorder (Sony MSW 2100 EP), and dirty or defective tapes are automatically cleaned.

2.2 Content-based video analysis

The aim of the content-based video analysis algorithms is to automatically assign semantic tags to videos for the purpose of facilitating content-based search and exploration. The fundamental problem is to overcome the discrepancy between the extracted features and the human interpretation of the (audio-)visual data. In the literature, this discrepancy is also known as "semantic gap". Smeulders et al. [37] describe the semantic gap as "the lack of coincidence between the information that one can extract from the visual data and the

[4] http://www.jordicom.ch/tv-media/.

[5] http://wdr-mediagroup.com.

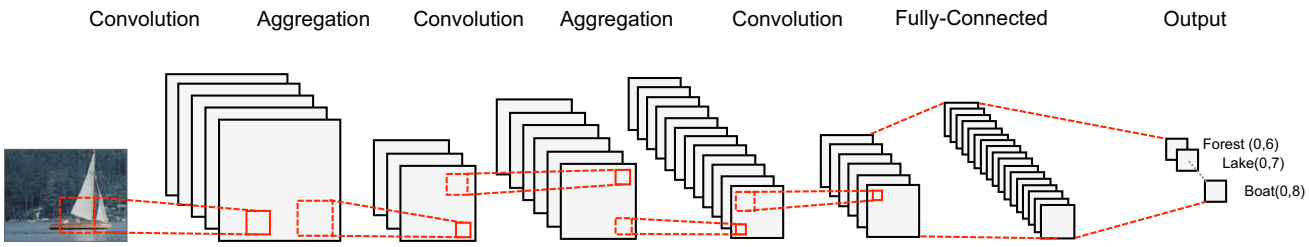

Fig. 2 Deep convolutional neural network

interpretation that the same data have for a user in a given situation".

Typically, automatically generated labels are assigned to video shots. Therefore, shot boundary detection has to be performed. The aim of shot boundary detection is the temporal segmentation of a video sequence into its fundamental units, the shots. A shot is generally understood as a video sequence recorded continuously without any interruption. The transitions between shots can be abrupt (cuts) or gradual (fade in/out, dissolves, wipes). Shots are detected using our shot boundary detection algorithms [10,11], some of which belonged to the top approaches at the TRECVid challenge 2007[6]. To detect gradual transitions more reliably, camera motion estimation (e.g., see [14]) is leveraged for false alarm removal.

Based on the results of the temporal video segmentation, concept classification, person and text recognition are applied for video analysis to automatically extract high-level content-based metadata. Furthermore, an index is generated for fast semantic similarity search in large video databases.

In the following sections, the content-based video analysis algorithms as well as the similarity search approach are described in more detail.

2.2.1 Visual concept classification

The classification of visual concepts is a challenging task due to the large complexity and variability of their appearance. Visual concepts can be, for example, objects, sites, scenes, personalities, events or activities. The definition of our GDR specific concept lexicon is based on the analysis of user search queries with a focus on queries that were experienced as difficult and time-consuming to answer manually. Considering the utility or usefulness for search queries, the observability by humans and the feasibility in the sense of automatic detection, a lexicon of 91 concepts was defined after analyzing more than 36,000 user queries received within a five-year period from 2008 to 2013. Therefore, user queries that are assumed to be of future research interest were summarized thematically and ordered by frequency. The concept lexicon comprises events such as *border control* and *concert*; scenes such as *railroad station* and *optical industry*; objects like *Trabant*; or activities such as *applauding*. To build the concept models, training data have to be annotated manually. For this purpose, a client–server-based annotation tool was built to facilitate the process of training data acquisition and to select a sufficient quantity of representative training examples for each concept.

Recently, deep learning algorithms fostered a renaissance of artificial neural networks, enabled by the massive parallel processing power of modern graphics cards. Deep learning approaches, especially deep CNNs, facilitated breakthroughs in many computer vision fields [4,17,22,36,39]. Instead of using handcrafted features such as SIFT descriptors [26], CNNs learn the features automatically during the training process. A CNN consists of several alternating convolution and aggregation (i.e., max-pooling) layers with increasingly complex feature representations and typically has several fully connected final layers, as shown in Fig. 2.

Most state-of-the-art network architectures for image recognition [18,22,38] as well as the current datasets [8,43] consider only a single concept per image ("single-label"). In contrast, real-world concept classification scenarios are multi-label problems. Several concepts, such as *summer*, *playground* and *teenager*, may occur simultaneously in an image or scene. While some approaches use special ranking loss layers [16], we extended the CNN architecture of the GoogleNet [38] using a sigmoid layer instead of the softmax layer in combination with a cross-entropy loss function.

Since the training of a deep CNN model from scratch requires millions of training images and due to the relatively small amount of available training data, we adapted a pre-trained CNN classification model (GoogleNet [38] trained on ImageNet [8]) to the new GDR concept lexicon using our multi-label CNN extension and performed a fine-tuning on the GDR television recordings. The models were trained and fine-tuned using the deep learning framework Caffe [20].

2.2.2 Similarity search

Since the DRA offers researchers a large number of video recordings containing several millions of video shots, the

[6] http://trecvid.nist.gov.

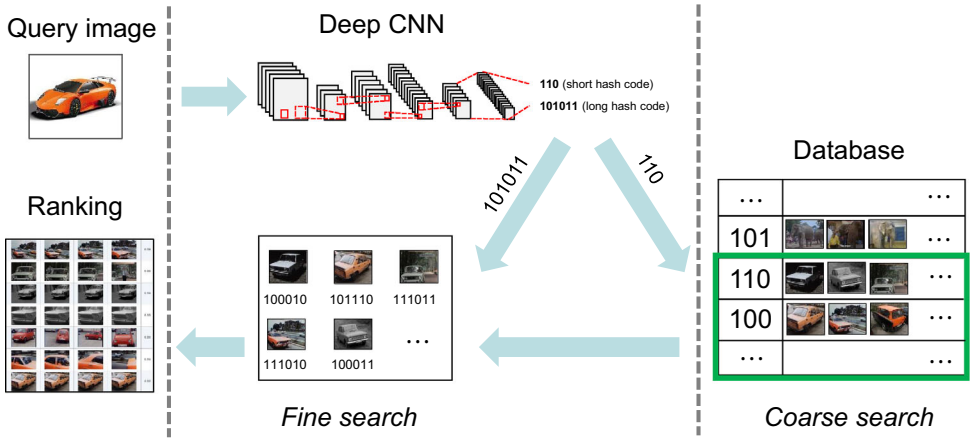

Fig. 3 Content-based similarity search

need for a system that helps to rapidly find desired video shots emerges. While scanning through the whole video archive is practically infeasible for humans, a possible solution is to index the videos via concepts as described in Sect. 2.2.1. However, this approach requires manually annotated training images for learning the concept models. Additionally, search queries are restricted to the vocabulary of predefined concepts and new concept models have to be developed on demand. In contrast to textual concept-based queries, image-based queries provide users more flexibility and a new way of searching.

While query-by-content based on low-level features turned out to be insufficient to search successfully in large-scale multimedia databases, image representations learned by deep neural networks greatly improved the performance of content-based image retrieval systems [41]. They are less dependent on pixel intensities and are clearly better suited for searching semantic content. However, high-dimensional CNN features are not well suited for searching efficiently in large video collections. Fast search in large databases is an essential requirement for practical use. For this purpose, proposals for learning binary image codes for compact representations and fast matching of images have been made. Krizhevsky and Hinton [21], for example, used deep autoencoders and Lin et al. [24] extended a CNN to learn binary hash codes for fast image retrieval.

In this section, an approach for fast content-based similarity search in large video databases is presented. To efficiently store and match images, the approach is based on binary image codes. The mapping of images to binary codes is often referred to as "semantic hashing" [35]. The idea is to learn a "semantic hash function" that maps similar images to similar binary codes. Our method extends the approach introduced by Lin et al. [24]. In our method, we fine-tune a VGG-16 CNN architecture [5], trained on the Places dataset [43], with an additional coding layer before the final classification layer. The learning rate is decreased following a polynomial decay with a power of 4. As an optimization method, we use SGD with a momentum of 0.9. In contrast to the approach of Lin et al. [24], we use hash codes for the refined search as well. In our two-stage approach, both coding layers, for 64-bit and 256-bit binary codes, are integrated into the same architecture and trained concurrently. Hence, sharing the CNN parameters for coarse and refined search completely eliminates the time for a fine-grained search formerly required for distance computations on high-dimensional float features at test time.

A further advantage of our approach is that the hash function can be adapted to unlabeled datasets by using the predictions of the pre-trained classification model for error propagation.

The overall system for similarity search is based on the analysis of keyframes, i.e., representative images. In our approach, five frames (the first, the last and three in between) per video shot are used as keyframes for indexing. Given the hash function, the keyframes of the video collection are fed into the deep CNN and the mapped binary codes are stored in the database. Based on the resulting index, queries-by-image can be answered by matching the binary code of the given image to the database. The overall retrieval process is shown in Fig. 3. Given the query image, the 64-bit and 256-bit binary codes are extracted using the learned deep CNN. First, a coarse search is performed using 64-bit binary codes, resulting in a comparatively short list of potential results. The Hamming distance is applied to compare the binary codes. A vantage point tree [42] is used as an additional index structure that recursively partitions the binary codes in the Hamming space into close and distant points by choosing so-called vantage points to significantly accelerate the nearest neighbor search. The resulting short list consists of 10,000 nearest neighbors. The longer the binary codes, the more accurate are the image representations. Therefore, in the second stage, a refined search using 256-bit binary codes is performed on the short list. The images are ranked according to the Hamming distance to the query image. Finally, the resulting images are mapped to video shots.

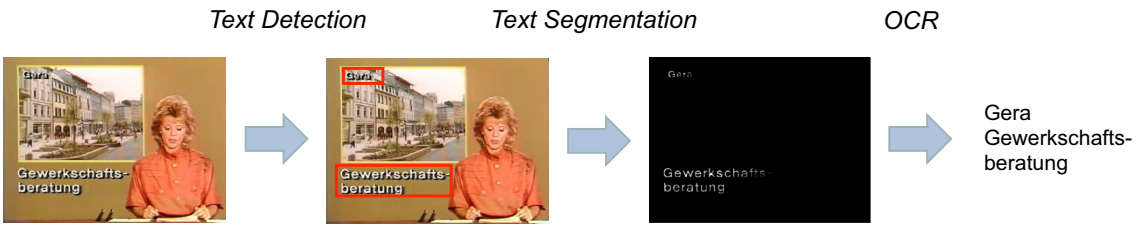

Fig. 4 Text recognition pipeline

2.2.3 Person recognition

Based on the analysis of user search queries, our GDR specific concept lexicon was extended by 9 personalities: *Erich Honecker*, *Walter Ulbricht*, *Hilde Benjamin*, *Siegmund Jähn*, *Hermann Henselmann*, *Christa Wolf*, *Werner Tübke*, *Stephan Hermlin*, and *Fritz Cremer*. Instead of using concept classification, persons are recognized using a face recognition approach [12]. For this purpose, feature representations of known persons are stored in a face database. A face recognition system was built that scans the video shots and recognizes the identity of a detected face image by comparing it to the face database. Finally, the resulting index of person occurrences can be used in search queries.

The face processing pipeline consists of several components: face detection, face alignment, and face recognition. For face detection, we used the method of Viola and Jones [40] due to its low runtimes. For face alignment and recognition, we used a commercial library, called FaceVACs[7], since it achieves better results on the GDR television recordings than approaches like Fisherfaces [3] or Local Binary Pattern Histograms [1] that are provided by the Open Computer Vision (OpenCV) Library[8]. Furthermore, we evaluated whether training data augmentation using Google search queries or a face tracking component would improve recognition accuracy. Since only a few training examples are provided for some of the personalities, using adding additional training images from the Internet led to a significant improvement in these cases. The face tracking component uses optical flow computation to assemble face appearances of the same person in subsequent frames [12]. The number of detection hits per face that is returned by the Viola and Jones face detector is used to select the best face within a face sequence with respect to size and frontal appearance. This face is then used for face recognition. The additional generation of face sequences did not lead to an improvement in the results and is thus not worth the extra computational effort. Since the training samples contain face images from a wide range of views, it does not seem to be beneficial to analyze only the most frontal face of a sequence.

2.2.4 Text recognition (video OCR)

Superimposed text often hints at the content of a video shot. In news videos, for example, the text is closely related to the current report. In silent movies, it is used to complement the screen action with cross-headings. The involved algorithms can be distinguished by their objectives, whether it is text detection, also called text localization, text segmentation, or optical character recognition [15] (see Fig. 4).

We have developed a text recognition system that allows users to search for in-scene and overlaid text within the video archive. For this purpose, the I-frames of the videos are analyzed, and the recognized ASCII text is stored in the database. Based on the resulting index, OCR search queries can be answered by a list of video shots ranked according to the similarity to the query term. Due to low technical video quality and low contrast of text appearances, the similarities between the query term and the words in the database are calculated using the Levenshtein distance [23].

For text detection, localization and segmentation in video frames, a method based on Maximally Stable Extremal Regions (MSER) was employed [28,33] using the OpenCV Library. It can detect both overlaid text and text within the scene, for example on banners. Experimental results have revealed that the text segmentation component plays an important role in the case of videos of low technical quality. Text segmentation crops the detected and as characters classified extremal regions out of the image to yield black letters on a white background. This step is necessary to feed the result into an OCR algorithm that transforms the image into machine-readable text. A non-uniform background would normally impair this process. For OCR, we evaluated two open source libraries: a Long Short-Term Memory network (LSTM) approach[9] [4] and Tesseract[10].

The OCR approaches were evaluated on about 11 h of video data including 14 videos of GDR television recordings from different years and genres. This subset was completely annotated with 3,781 words referring to video shots. As our evaluation metric, we used the accuracy score that we calculated as the number of correct recognized characters divided

[7] http://www.cognitec.com.

[8] http://opencv.org.

[9] https://github.com/tmbdev/ocropy.

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

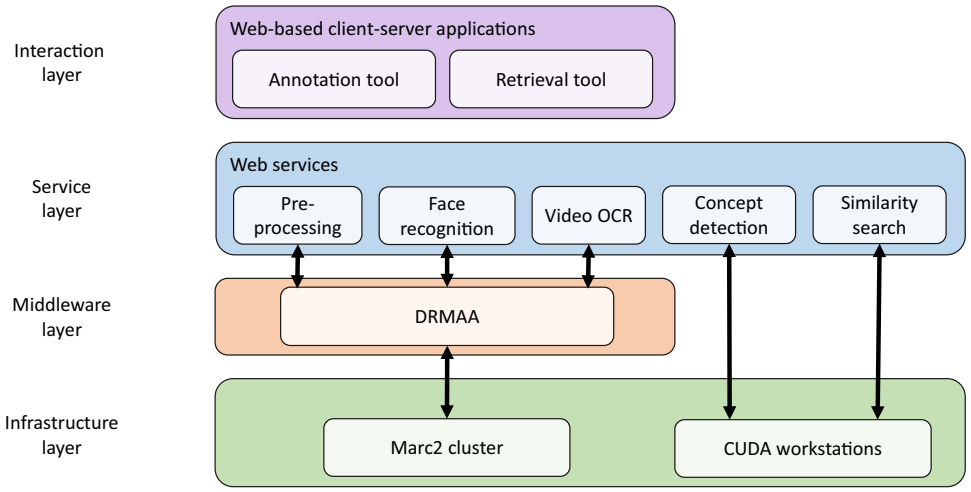

Fig. 5 Service-oriented architecture for CBVR

by the total number of characters. The number of correct characters per word was calculated as follows:

$$\max(\text{length}(w_{\text{rec}}), \text{length}(w_{\text{gt}})) - d_L(w_{\text{rec}}, w_{\text{gt}}) \quad (1)$$

where w_{rec} is the recognized word, w_{gt} is the ground-truth and d_L is the Levenshtein distance. Based on this quite challenging subset of GDR television recordings, Tesseract achieved an accuracy of 40% in comparison with the LSTM-based approach that achieved 26%. Thus, we used the Tesseract approach in Sect. 3.

2.3 Service-oriented architecture

A service-oriented architecture is used to deal with the requirements of the content-based video retrieval system. Due to the large amount of video data and the computationally expensive video analysis algorithms, a distributed heterogeneous architecture is employed to provide scalability. An overview of the architecture is given in Fig. 5. User interaction is handled via web-based client-server applications providing interfaces to use the following web services: preprocessing, person recognition, video OCR, concept classification and similarity search. The GUIs of the annotation and retrieval tool are described in more detail in Sects. 2.4 and 3.

The web services are executed on different hardware architectures. Here, we have to distinguish between CPU and GPU algorithms. While face recognition, video OCR and the preprocessing steps including shot boundary detection are CPU intensive, concept classification and similarity search mainly use GPU resources.

For the CPU-intensive algorithms, the MARC2 computing cluster at the University of Marburg, Germany, was used. The MARC2 cluster has 96 compute nodes, each consisting of 4 AMD Opteron 6276 or 6376 with 16 cores@2.3 GHz each, i.e., 6144 cores. In addition, there are two head nodes with the same specification. In total, MARC2 has 24 TB RAM and 192 TB of disk storage space. The operating system is Red Hat Enterprise Linux for the headnodes and CentOS for the compute nodes. The Sun Grid Engine 6.2u5 (SGE) is used as the job scheduler. For this purpose, specific interfaces were defined. The CPU-intensive algorithms were ported to the MARC2 infrastructure and were encapsulated in separate jobs. Web services for preprocessing, face recognition, and video OCR are provided. The preprocessing web service starts several jobs for transcoding the videos, for shot boundary detection and for extracting images and thumbs. The web services submit jobs to the SGE, control job execution and provide status information. For management purposes, the distributed resource management application API (DRMAA) is used. Instead of transferring user data directly, only references are sent via parameters during the service call. The images and videos are stored on a data server and the actual data transport is handled via network file system shares. The advantage is an overlap of data transfer and service execution, which contributes to the improvement in the overall runtime performance.

The web services for concept classification and similarity search use GPU resources and are installed on dedicated servers with Nvidia Geforce GTX 770 graphics cards with 4GB of memory. Two similarity search-related web services exist. The first one is responsible for hash code generation from videos, for uploading the extracted codes to the database and for updating the index structures. The second one provides similarity search, taking a query image as input and returning a sorted list of the most similar video shots.

While the execution of models for concept classification and similarity search is really fast on GPUs (only a few milliseconds), training of such deep CNNs is computationally expensive even on graphics cards. Considering the hardware requirements concerning processing power, GPU memory, main memory and hard disk capacities, we have built a highly

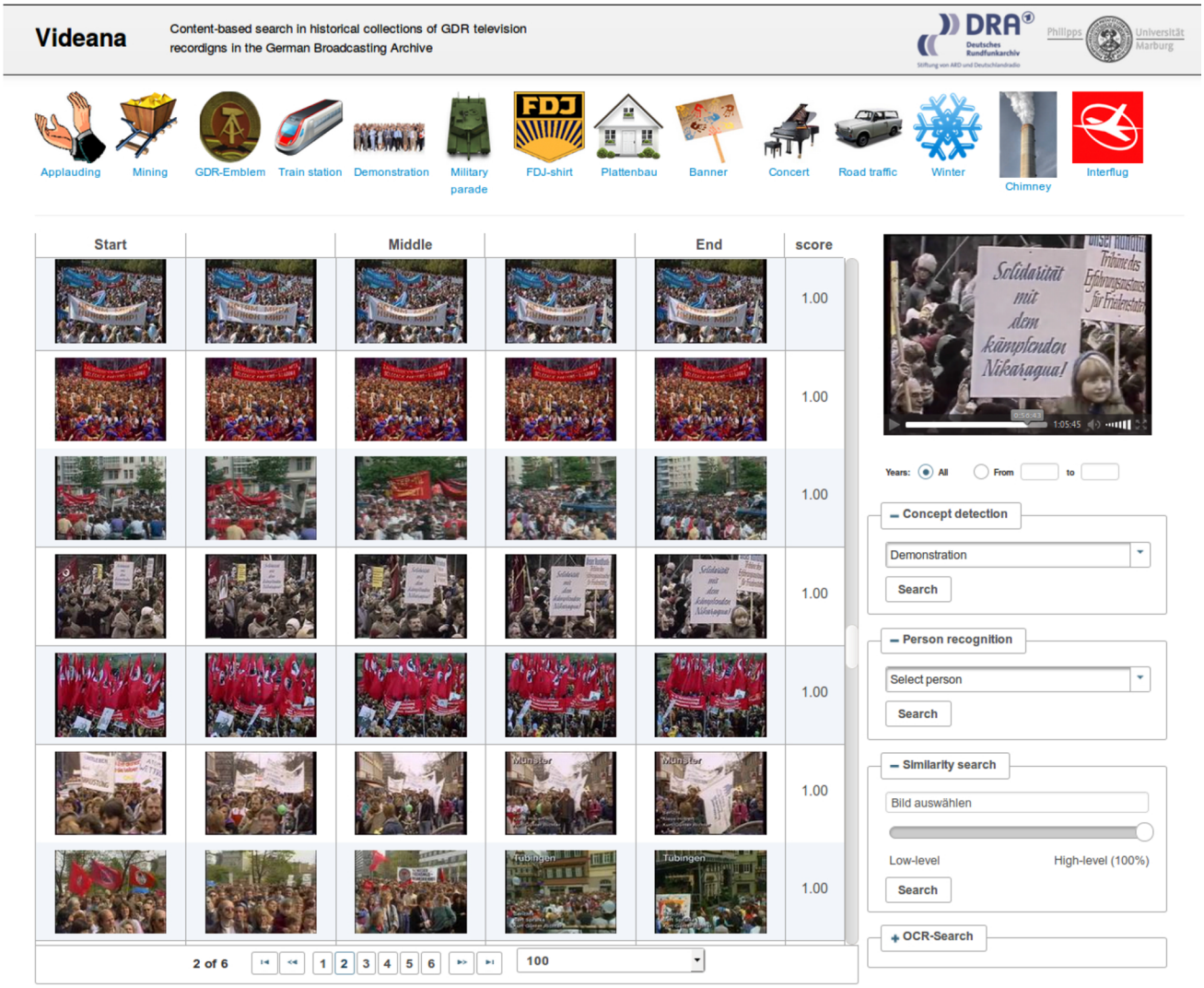

Fig. 6 GUI of the retrieval tool showing the results for the concept *demonstration*

optimized system for deep learning similar to the Nvidia DevBox[11] to train deep neural network models. The system consists of four GeForce GTX Titan X GPUs with 12GB RAM and 3072 CUDA cores at 1000/1075 MHz, an Intel Core i7-5930K CPU with six cores at 3.50GHz, 64 GB of DDR-4 RAM, 8 TB of disk space for large datasets and a 250 GB SSD for fast I/O operations.

2.4 Workflows

The client–server-based web applications, the annotation and the retrieval tool provide simple interfaces to perform content-based video analysis and retrieval in a distributed, heterogeneous environment.

The retrieval tool provides image and scene search in the metadata-enriched video collection. A web-based GUI was developed to automatically respond to user queries related to concepts, persons, similar images, or text. The retrieval results are presented to the user in the form of a ranked list of video shots (see Fig. 6) where each video shot is represented by five key frames and a probability score indicating the relevance of the shot. Furthermore, a video player allows to visually inspect the video shots.

Administrative tasks, video uploads, manual annotations and orchestration of the content-based video analysis jobs are handled via the annotation tool. It provides an upload page where videos can be selected and uploaded to the system. The video upload automatically triggers the preprocessing jobs for shot boundary detection and visualization purposes by considering algorithm and data parallelism to run as many processes as possible concurrently. The preprocessing work-

[11] https://developer.nvidia.com/devbox.

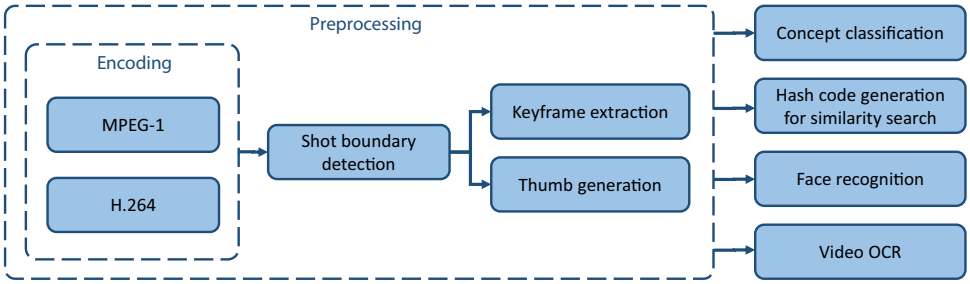

Fig. 7 Web service dependencies and workflows

flow consists of the following jobs: video transcoding, shot boundary detection, keyframe extraction and thumbnail generation (see Fig. 7). The MPEG-1 encoding is due to the use of our own shot boundary detection algorithm that directly uses DC components and motion vectors of the MPEG-1 video format. We are currently working on an integration of the ffmpeg interface to omit the additional step of MPEG-1 encoding. While MPEG-1 videos are necessary as a standardized input for the shot boundary detection algorithm, the MP4 videos are compressed using the H.264 codec to generate small videos suitable for web applications. The keyframes are extracted for the subsequent content-based analysis jobs, and the thumbnails are generated for the visualization of shots.

After the videos have been uploaded and preprocessed, they can be manually annotated for training data acquisition based on the predefined lexicon of visual concepts. The application provides GUIs for shot-based and interval-based labeling. The GUI for annotating intervals mainly consists of a video player and a video slider where concept occurrences can be inserted, deleted and edited. The interval-based labeling is more accurate than the shot-based labeling, but also more time-consuming.

As soon as the models for concept classification and person recognition have been built and installed, the job management page of the annotation tool can be used to start content-based analysis jobs to generate metadata for concepts, persons and text as well as hash codes. The dependencies of the different web services and jobs are visualized in Fig. 7.

3 Experimental results

Several experiments were performed. First, the performance of the deep multi-label CNN-based approach for concept classification was compared to a BoVW approach, motivating the use of CNN-based algorithms in the final system. Second, the content-based video retrieval system was evaluated on historical GDR television recordings.

The results are evaluated using the average precision (AP) score:

$$\mathrm{AP}(\rho) = \frac{1}{\left|R \cap \rho^N\right|} \sum_{k=1}^{N} \frac{\left|R \cap \rho^k\right|}{k} \psi(i_k) \tag{2}$$

$$\text{with} \quad \psi(i_k) = \begin{cases} 1 & \text{if } i_k \in R \\ 0 & \text{otherwise} \end{cases}$$

where N is the length of the ranked shot list, $\rho^k = \{i_1, i_2, \ldots, i_k\}$ is the ranked shot list up to rank k, R is the set of relevant documents, $\left|R \cap \rho^k\right|$ is the number of relevant video shots in the top-k of ρ and $\psi(i_k)$ is the relevance function. Generally speaking, AP is the average of the precisions at each relevant video shot. To evaluate the overall performance, the mean AP score is calculated by taking the mean value of the AP scores from different queries.

3.1 BoVW versus deep multi-label CNN

To investigate the performance of the proposed deep multi-label CNN for concept classification, a comparison between a BoVW approach based on state-of-the-art handcrafted features and the deep CNN-based approach was performed on the fully annotated, publicly available NUS-WIDE scene dataset [7]. The NUS-WIDE scene subset covers 33 scene concepts and consists of 34,926 images in total. Half of the images are used as the training set and the rest as the test set. For the deep multi-label CNN classifier, a GoogLeNet pre-trained on the ILSVRC 2012 dataset was fine-tuned on the NUS-WIDE scene training images.

Using the BoVW approach, an image or a video shot is represented as a histogram of visual words by mapping the local descriptors to a precalculated visual codebook. The BoVW approach is based on a combination of different feature representations relying on optimized SIFT variants. These SIFT variants use different sampling strategies: a dense sampling strategy and a Difference of Gaussians (DoG) keypoint detector [26]. Color information is integrated using concatenated SIFT descriptors from different color channels (transformed color SIFT, RGB-SIFT) and a combination of SIFT descriptors and local color moments. Furthermore, spatial information is captured using a spatial pyramid of 1×1 and 2×2 equally sized subregions. Altogether, four different SIFT variants are used.

The visual codebooks are generated using a K-means algorithm and consist of 5000 visual words. Instead of mapping a SIFT descriptor only to its nearest neighbor or to all visual words, the codebook candidates are locally constrained to the

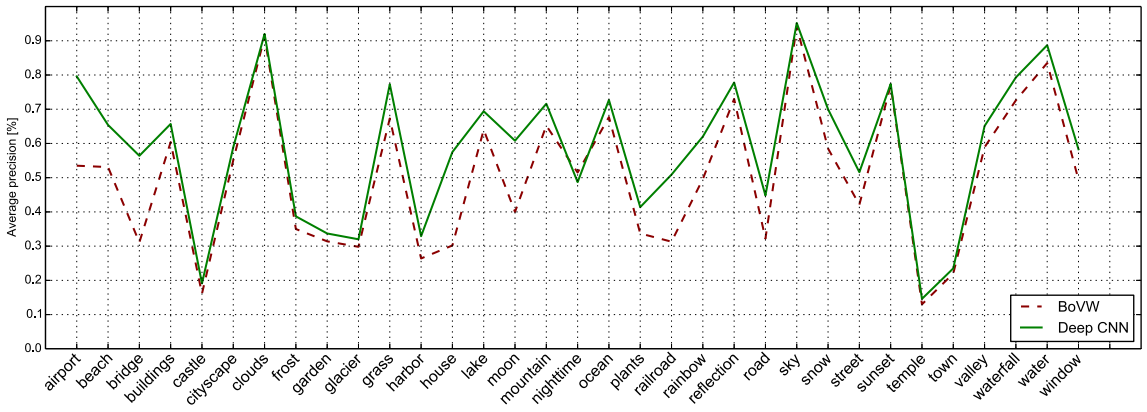

Fig. 8 Performance comparison between the BoVW and the deep multi-label CNN approach on the NUS-WIDE scene dataset

five nearest visual words. This locality constraint was shown to be superior for BoVW approaches [25].

The different feature representations are combined in a support vector machine (SVM) classifier using multiple kernel learning [9,13,29–31]. For all feature representations, the χ^2-kernel is used to measure the similarities between the data instances.

Figure 8 shows the comparison between the BoVW and the multi-label deep CNN approach on the NUS-WIDE scene dataset for each concept. While the BoVW approach achieves a mean AP of 50.3%, the deep CNN approach obtains 58.6%. Thus, the novel deep multi-label CNN approach significantly outperforms the BoVW approach on the NUS-WIDE scene dataset by a relative performance improvement in almost 20%. Results on the other NUS-WIDE subsets are 56.3 and 79.62% on NUS-WIDE objects and 40.83 and 55.24% on NUS-WIDE lite for BoVW and CNN, respectively.

In contrast to binary SVM classifiers, deep neural networks are inherently capable of processing multiple classes, such that only a single compact model has to be built for all concept classes. While the runtime of the BoVW approach is already 1.97 s for feature extraction on the CPU and the classification runtime depends linearly on the number of concepts, the multi-label CNN takes less than a second on the CPU (Intel Core i5) and is even considerably faster on the GPU.

Although deep neural networks are computationally expensive in the training phase, they are very efficient during classification. Altogether, multi-label CNNs provide clearly better recognition quality, compact models and significantly faster classification runtimes.

3.2 Historical GDR television recordings

In this section, the content-based video retrieval algorithms for concept classification, similarity search, person recognition and video ocr are quantitatively and qualitatively evaluated based on the video collection of historical GDR television recordings. In total, more than 3000 h of historical GDR television recordings were digitized. The video footage is technically quite challenging. Many recordings are in grayscale and of low technical quality; the older the recordings, the poorer the video quality. The temporal segmentation of the videos resulted in approximately 2 million video shots. From these shots, 416,249 were used for the training process and 1,545,600 video shots, corresponding to about 2500 h of video data, for testing.

The developed retrieval tool provides a web-based GUI to submit user queries related to concepts, persons, similar images and text. The retrieval results are presented to the user in the form of a ranked list of video shots (see Fig. 6) where each video shot is represented by five key frames and a probability score indicating its relevance. Furthermore, a video player allows to visually inspect the video shots. In the following, the results for concepts and persons as well as sample queries for similar images and text are presented.

3.2.1 Concept classification and person recognition

In total, 86 concepts, consisting of 77 concepts and 9 persons, were evaluated.

From the original 91 concepts, 14 were dismissed due to an insufficient number of training images. These omitted concepts, including rarely occurring concepts with non-distinctive training samples like “Kinderkaufhaus Straußberger Platz”, “Werner Seelenbinder Halle” or “Hotel Unter den Linden”, had less than 75 positive images sampled only from a few scenes. An overview of the evaluated concepts and the corresponding number of positive training samples is given in Table 1. Although a correlation between the AP score and the number of positive training samples is observable, the comparison of rarely and frequently occurring concepts in terms of AP is difficult since the AP score depends on the frequency of a concept class in the test set.

Table 1 Concept detection results in terms of average precision

Concept	AP		#
	Top 100 (%)	Top 200 (%)	
GDR emblem	100.0	100.0	2157
City	100.0	100.0	5121
Black-and-white	100.0	100.0	1346
Man	100.0	99.8	5000
Leipziger Messe	100.0	99.5	1548
Summer	100.0	99.5	2300
Demonstration/rally	100.0	99.3%	2512
FDJ shirt	99.9	99.1	3757
Plattenbau	99.7	97.9	3307
Tram	98.9	96.8	1930
Animal husbandry	98.8	96.3	2204
Military parade	98.5	95.9	2809
Airport	98.4	94.5	1955
Photo	98.2	98.1	1731
Woman	98.2	97.5	5000
Road traffic	95.8	93.8	3032
Pioneer	95.5	90.5	2375
Concert	94.0	88.5	2595
Shake hands	92.9	91.6	1419
Winter	92.6	89.5	5302
Banner/slogan	90.5	92.2	2581
Interflug (logo)	88.0	73.6	669
Kitchen	87.5	81.8	2430
Steel factory	86.9	83.1	2504
Daylight mining	84.3	78.1	2759
Textile factory	83.6	76.9	2209
Railroad station	83.1	73.0	1291
Chimney	82.3	78.4	959
Apartment construction	82.2	77.6	2753
Ambulance	82.2	73.5	448
Dance	81.4	70.4	1667
Production hall	80.8	79.4	1879
Village	80.6	76.5	1607
Deutsche Reichsbahn	80.3	77.1	2384
People in pedestrian zone	79.1	74.4	1192
GDR flag	78.6	71.8	1056
Kaufhalle (supermarket)	77.7	66.3	1622
ABV (community policeman)	75.4	67.3	1929
Wreath ceremony	73.4	68.4	1585
Teenager	69.1	67.2	2131
Camping site	67.2	55.8	98
Non-renovated apartment building	67.1	59.4	835
Beach scene	66.7	56.5	821
Trabant	66.1	60.1	2201
Church outside	64.9	60.8	876
Church inside	63.5	50.0	603
Applause	62.2	56.4	2333
Allotment	61.2	54.2	1099
Food industry	60.7	49.7	1082
Playground	58.3	47.6	632
Optical industry	52.8	46.3	520
Kindergarten	46.2	35.3	1854
State council building	45.0	34.6	1156
Simson (logo)	36.6	32.7	90
Swimming pool	33.3	30.7	759
Nurse	32.8	28.9	1164
Launching	32.6	26.8	96
Microelectronics	32.2	12.9	1281
Automobile industry	29.5	26.5	617
Mining	27.6	20.4	1885
Kiosk	26.2	21.6	80
Disco	23.9	19.7	611
Border control	22.6	19.1	89
Dockyard	22.0	21.7	1079
Theater performance	21.1	19.4	1172
Priest	13.9	13.9	91
Berlin Wall	13.7	12.1	88
Narrow-gauge railway	13.5	12.3	444
Power station	12.4	11.9	603
Prison	8.9	7.1	87
Brotherly kiss	6.8	4.3	114
Charité	6.6	6.6	75
SERO	5.9	5.9	99
Church service	5.4	5.0	85
Double-decker bus	5.2	4.8	93
Swastika (logo)	2.8	2.3	76
Waiting queue	1.0	1.4	77
	62.4	58.0	118,020

The last column contains the number of positive training images

It can be easily shown that the AP score of a randomly generated retrieval result tends toward the concept frequency in the given test set.

Altogether, 118,020 positive training examples were gathered for learning the concept model. The retrieval results for concepts and persons were evaluated based on the top-100 and top-200 ranked video shots. Although 14 concepts have less than 100 training images, and despite poor video quality, we obtained mean AP scores of 62.4 and 58.0% for the top-100 and top-200, respectively. Even concepts occurring predominantly in grayscale shots of low video quality yielded good results, such as *daylight mining* with 84.3% AP. These results reveal the high robustness of the proposed multi-label

Table 2 Face recognition results in terms of average precision

Person	Top 100 (%)	Top 200 (%)
Erich Honecker	100	100
Walter Ulbricht	100	100
Hilde Benjamin	98.6	96.2
Siegmund Jähn	98	98
Hermann Henselmann	85.6	85.7
Christa Wolf	76.4	76.4
Werner Tübke	65	65
Stephan Hermlin	64.3	47.1
Fritz Cremer	61.6	61.6

deep CNN approach with respect to the low quality historical video data. Since users of the German Broadcasting Archive are often looking for everyday scenes in the former GDR, concepts such as *pedestrian*, *supermarket*, *kitchen*, *camping site*, *allotment* or *production hall*, are valuable contributions to help researchers in finding appropriate scenes. The system was extensively tested by an archivist in the everyday work of the German Broadcasting Archive. From this archivist's perspective, it has been shown that concepts with an AP score of more than approximately 50% turned out to be very useful in practice. Altogether, 66% of the concepts achieved an AP score of more than 50%.

Furthermore, searching manually for persons in videos is a quite time-consuming task, particularly for more rarely occurring persons or less known members of the *Politbüro* and Ministers of the GDR. For person recognition, we achieved a very good result of 83.3% mean AP on the top-100 and 81.1% mean AP on the top-200 video shots. For distinctive and frequently occurring personalities, such as Erich Honecker and Walter Ulbricht, an AP score of 100% was achieved, and even for more rarely occurring personalities, such as Werner Tübke, Stephan Hermlin or Fritz Cremer, the AP score on the top-100 video shots is over 60%, as shown in Table 2. Thus, the high quality of the provided automatic person indexing algorithms is a great benefit for archivists as well as for users of the archive.

In total, we achieved a mean AP of 64.6 and 60.5% on the top-100 and top-200, respectively, for both concepts and persons.

3.2.2 Similarity search

The interpretation whether two images are similar is subjective and context specific. The definition of similarity ranges from pixel-based similarity to image similarity based on the semantic content. How much low-level and semantic similarity contribute to the retrieval results can be individually adjusted in the GUI. Furthermore, two use cases were implemented: searching by video frames selected from the corpus and searching by external images, e.g., downloaded from the Internet. In our evaluation, we focus on the more difficult task of semantic similarity using 50 external query images from the Internet chosen collaboratively by computer scientists and archivists. Each retrieval result was evaluated up to the first 100 video shots. Altogether, two systems were evaluated using an image corpus of more than 7 million keyframes. Both systems use the same 256-bit binary codes. While the baseline system is a one-stage approach using only 256-bit binary codes our two-stage approach uses an additional vantage point tree based on 64-bit binary codes and the 256-bit codes for the fine search. The results are presented in Table 3. In contrast to the one-time training process which takes a few days for building the deep CNN model, the overall response time for a query is much more important. While the overall runtime of the baseline system is almost 30 s for a query, we achieve a very fast response time of less than 2 s on the average for a similarity search query using our two-stage approach in combination with a vantage point tree. Even the performance in terms of mean AP is slightly better than the baseline system, by obtaining a mean AP of 57.5%. The AP results of the individual search queries are shown in Fig. 9.

Table 3 The similarity search results are based on 50 query images chosen from the Internet and are evaluated in terms of mean average precision based on the top-100 video shots

	Average runtime per query (s)	Mean AP top 100 (%)
Baseline system	29.53	57.0
Our approach	1.99	57.5

The runtimes were measured on a server with a Nvidia Geforce GTX 770 graphics card and an Intel Core i7-6700K CPU with eight cores at 4.00GHz

An example result for a query image showing a meal is presented in Fig. 10. More retrieval results are visualized in Fig. 11 where the first column shows the query images downloaded from the Internet followed by the first six highest ranked keyframe images. Although the results of the last two rows are somehow similar, they do not show the shape the archivist requested. Only the sixth result image of the last two rows is similar in terms of shape.

To summarize, the implemented similarity search system significantly extends the accessibility to the data in a flexible way. It provides complementary search queries that are often hard to verbalize. In addition, it facilitates incremental search. Previous results may serve as a source of inspiration for new similarity search queries for refining search intentions.

3.2.3 Video OCR

Another useful search option is offered by video OCR. OCR search results are very helpful since overlaid text

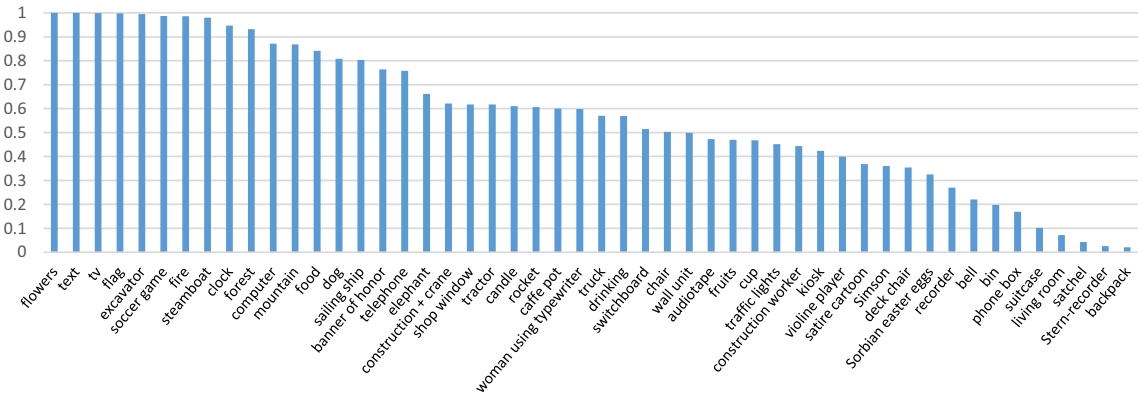

Fig. 9 Similarity search results for 50 query images from the Internet in terms of average precision evaluated up to the first 100 video shots

Example: A user is searching for material for the film production *Flavors in the GDR* and uploads an image of a meal. By using this query image, the user can carry out the search without words or meta-data while the ranked results contain a large number of relevant shots.

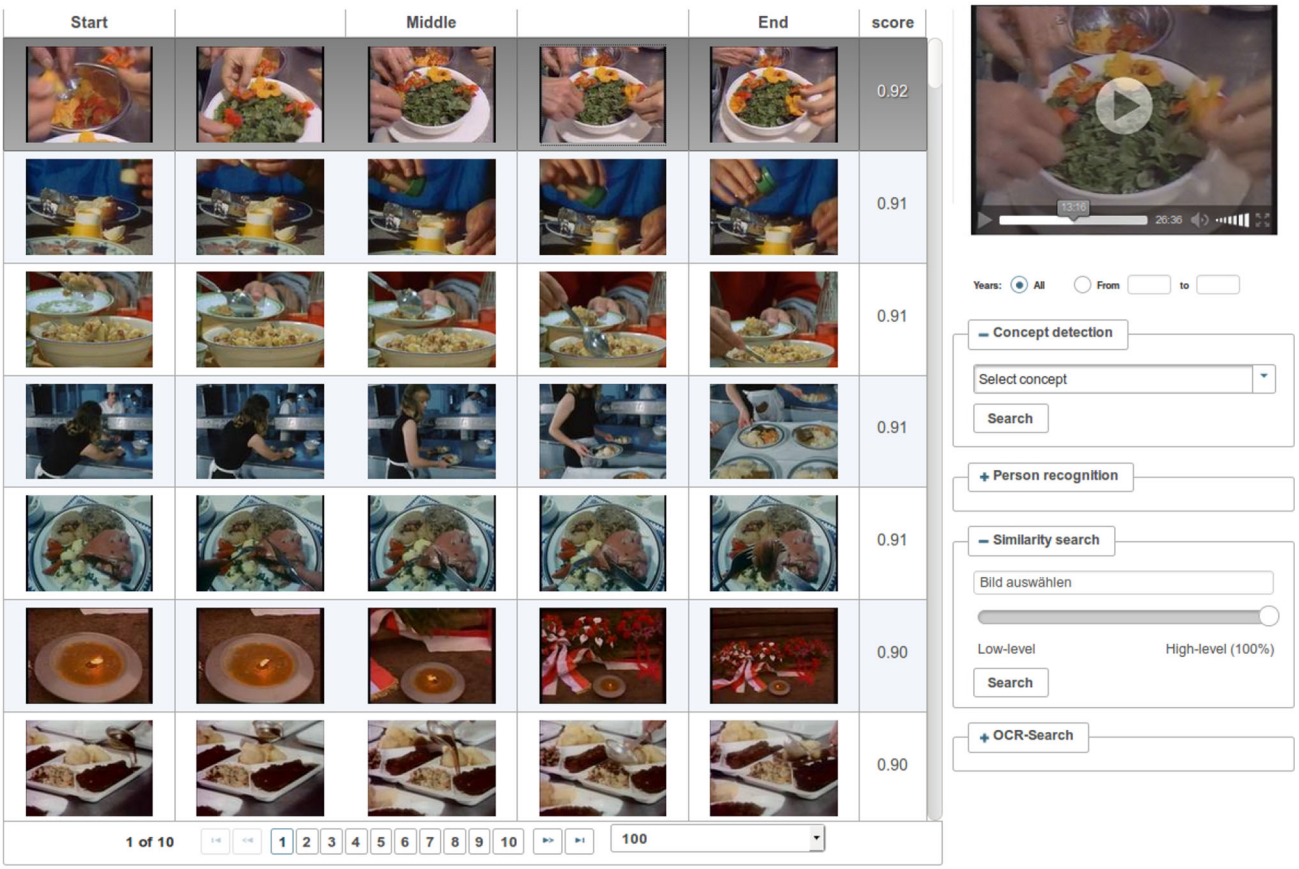

Fig. 10 A similarity search result for a query image showing a meal

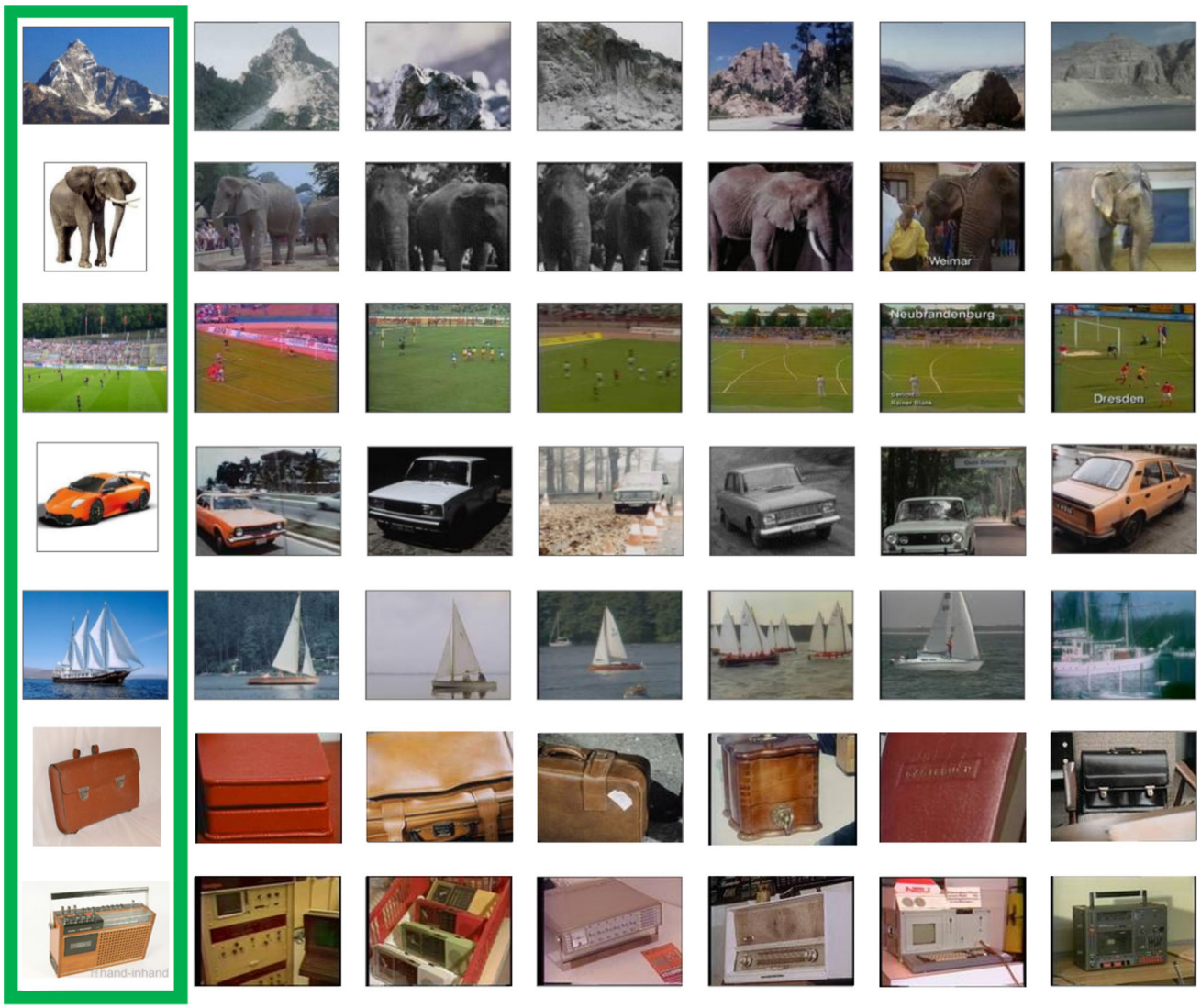

Fig. 11 Retrieval results for query images downloaded from the Internet. The first column shows the query images followed by first six highest ranked images from the database

is often closely related to the video content. For the task of text retrieval, 46 query terms according to previously observed search query preferences of DRA users were evaluated. Based on these 46 search queries, like *Abschaffung der Todesstrafe* (abolishment of death penalty), *Mikroelektronik* (microelectronics), *Öffnungszeiten* (opening hours), *Protestbewegung* (protest movement), *Rauchen verboten* (no smoking), *Warschauer Vertrag* (Treaty of Warsaw), *Planerfüllung* (plan fulfillment), *Gleichberechtigung* (equal rights), *Nationale Front* (national front), *Staatshaushalt* (national finances), or *Kinder- und Jugendspartakiade* (children and youth spartakiad), a very satisfying retrieval performance of 92.9% mean AP has been obtained. The average precision results for the 46 queries are shown in Fig. 12. The system retrieves the majority of slogans, locations, and other terms correctly. As expected, the results for overlaid text are significantly better than for text within the scene.

3.2.4 Use case: querying a TV production

In the following, the benefits of the content-based video retrieval system are illustrated based on search requests that arose in the context of the TV production *Das Erbe der Nazis* (heritage of the Nazis). While these queries are manually difficult and time-consuming to answer, the content-based video retrieval system is able to yield fast and practically useful results. Example queries are:

– *Everyday scenes in the GDR across the decades* 1950*s*/60*s*/70*s*/80*s*:

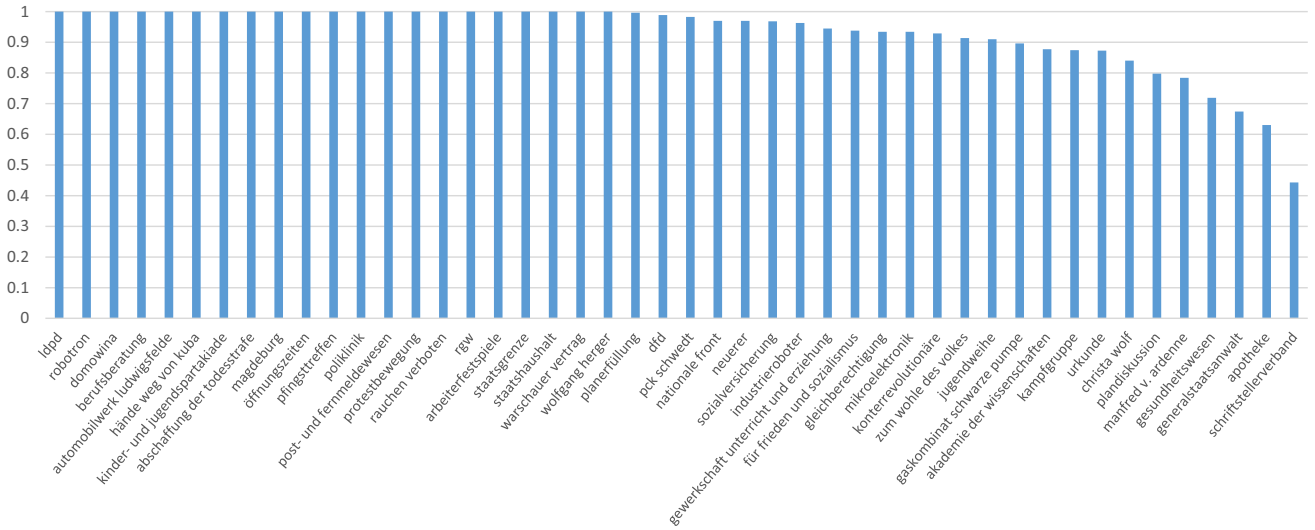

Fig. 12 OCR retrieval results for 46 text queries in terms of average precision evaluated up to the first 100 video shots

Using the content-based video retrieval system, this query can be answered by specifying the years and searching for concepts concerning everyday life in the GDR, e.g., *pedestrian* or *kitchen*. These results may serve as a starting point for further similarity search queries.

- *Monday demonstrations in 1989, crowds, banners "Wir sind das Volk" ("We are the people")*:
  The retrieval result for this query can be obtained by combining queries for the concepts *demonstration* and *banners/slogans*, an OCR query for in-scene text and a restriction to the years 1989.
- *Erich Honecker, parade, 1970s*:
  Useful results for this query can be found by combining a query for the person *Erich Honecker*, a query for the concept *military parade* and restricting the years to the 70s.

To summarize, the fine-grained automatic annotation is a very valuable supplement to human-generated metadata. Due to the variability of the content-based video retrieval system, different user needs are taken into account. The combination of different search modalities allows the DRA to answer a wide range of user queries leading to more precise results in significantly less time.

4 Conclusion

The DRA maintains the cultural heritage of television of the former GDR. The uniqueness and importance of this archive causes a great interest in the video content. In this paper, we have presented a novel content-based video retrieval system for searching in historical collections of GDR television recordings. It includes digitization, a service-oriented architecture for large-scale video analysis, a novel concept lexicon and extended state-of-the-art algorithms for visual concept classification, similarity search, person recognition, and video OCR. The content-based video retrieval system complements human annotations and supports archivists and users, respectively, in finding relevant video shots. Experimental results on about 2500 h of GDR television recordings have shown the usefulness of the video retrieval system.

There are several areas for future work. First, the concept-based approach requires manually annotated training images for learning the concept models. Since cultural research interests and topics may change in the future, the concept lexicon has to be continuously updated. New concepts have to be developed on demand. Therefore, it is important to reduce the effort for training data acquisition. Second, the audio modality can be used to improve the detection performance of several audio-related concepts [32]. Finally, a similarity search query can be quite subjective depending on a specific user in given situation. New strategies have to be developed to predict a user's intention.

Acknowledgements This work is financially supported by the German Research Foundation (DFG; Funding Programme: "Förderung herausragender Forschungsbibliotheken"; Project: "Bild- und Szenenrecherche in historischen Beständen des DDR-Fernsehens im Deutschen Rundfunkarchiv durch automatische inhaltsbasierte Videoanalyse"; CR 456/1-1, EW 134/1-1, FR 791/12-1).

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face recognition with local binary patterns. In: Proceedings of the IEEE European Conference on Computer Vision. pp. 469–481 (2004)

2. Albertson, D., Ju, B.: Design criteria for video digital libraries: categories of important features emerging from users' responses. Online Inf. Rev. **39**(2), 214–228 (2015)
3. Belhumeur, P.N., Kriegman, D.J.: Eigenfaces versus fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997)
4. Breuel, T.M., Ul-Hasan, A., Al-Azawi, M.A., Shafait, F.: High-performance OCR for printed English and Fraktur using LSTM networks. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 683–687 (2013)
5. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of the British Machine Vision Conference, pp. 1–11 (2014)
6. Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., Wactlar, H.: Informedia digital video library. Commun. ACM **38**(4), 57–58 (1995)
7. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: A real-world web image database from National University of Singapore. In: Proceedings of the ACM International Conference on Image and Video Retrieval, pp. 48:1–48:9 (2009)
8. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09), pp. 2–9 (2009)
9. Ewerth, R., Ballafkir, K., Mühling, M., Seiler, D., Freisleben, B.: Long-term incremental web-supervised learning of visual concepts via random savannas. IEEE Trans. Multimed. **14**(4), 1008–1020 (2012)
10. Ewerth, R., Freisleben, B.: Video cut detection without thresholds. In: Proceedings of the 11th International Workshop on Signals, Systems and Image Processing (IWSSIP '04), pp. 227–230. Poznan, Poland (2004)
11. Ewerth, R., Freisleben, B.: Unsupervised detection of gradual video shot changes with motion-based false alarm removal. In: Proceedings of the 11th Conference on Advanced Concepts for Intelligent Vision Systems, pp. 253–264 (2009)
12. Ewerth, R., Mühling, M., Freisleben, B.: Self-supervised learning of face appearances in TV casts and movies. Int. J. Semant. Comput. **1**(2), 185–204 (2007)
13. Ewerth, R., Mühling, M., Freisleben, B.: Robust video content analysis via transductive learning. ACM Trans. Intell. Syst. Technol. (TIST) **3**(3), 1–26 (2011)
14. Ewerth, R., Schwalb, M., Tessmann, P., Freisleben, B.: Segmenting Moving Objects in MPEG Videos in the Presence of Camera Motion. In: Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on IEEE, pp. 819–824 (2007)
15. Gllavata, J., Ewerth, R.: Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. In: Proceedings of 17th International Conference on Pattern Recognition (ICPR '04), pp. 425–428. IEEE (2004)
16. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep Convolutional Ranking for Multilabel Image Annotation. arXiv preprint arXiv:1312.4894 (2013)
17. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), pp. 6645–6649 (2013)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
19. Hentschel, C., Blümel, I., Sack, H.: Automatic annotation of scientific video material based on visual concept detection. In: Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies, p. 16 (2013)
20. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678 (2014)
21. Krizhevsky, A., Hinton, G.: Using very deep Autoencoders for content-based image retrieval. In: Proceedings of the European Symposium on Artificial Neural Networks, pp. 1–7 (2011)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1–9 (2012)
23. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Doklady. **10**, 707–710 (1966)
24. Lin, K., Yang, H., Hsiao, J., Chen, C.: Deep learning of binary hash codes for fast image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 27–35 (2015)
25. Liu, L., Wang, L., Liu, X.: In defense of soft-assignment coding. In: Proceedings of the 13th IEEE International Conference on Computer Vision, pp. 2486–2493 (2011)
26. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
27. Marchionini, G., Geisler, G.: The open video digital library. D-Lib. Mag. **8**(12), 1082–9873 (2002)
28. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image Vis. Comput. **22**(10), 761–767 (2004)
29. Mühling, M., Markus, M., Ewerth, R., Freisleben, B.: Improving cross-domain concept detection via object-based features. In: Proceedings of the International Conference on Computer Analysis of Images and Patterns (CAIP '15) (2015)
30. Mühling, M., Ewerth, R., Freisleben, B.: On the spatial extents of SIFT descriptors for visual concept detection. In: Proceedings of the 8th International Conference on Computer Vision Systems, pp. 71–80. Springer (2011)
31. Mühling, M., Ewerth, R., Shi, B., Freisleben, B.: Multi-class object detection with hough forests using local histograms of visual words. In: Proceedings of 14th International Conference on Computer Analysis of Images and Patterns, pp. 386–393. Springer (2011)
32. Mühling, M., Ewerth, R., Zhou, J., Freisleben, B.: Multimodal video concept detection via bag of auditory words and multiple kernel learning. In: Proceedings of the 18th International Conference on Advances in Multimedia Modeling, pp. 40–50. Springer (2012)
33. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2012)
34. Sack, H., Plank, M.: AV-Portal: The German National Library of Science and Technology's Semantic Video Portal. ERCIM News 96 (2014)
35. Salakhutdinov, R., Hinton, G.: Semantic hashing. Int. J. Approx. Reason. **50**(7), 969–978 (2009)
36. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
37. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern. Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
38. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
39. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: Closing the gap to human-level performance in face verification. In: Pro-

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2014)
40. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 511–518 (2001)
41. Wan, J., Wang, D., Hoi, S.C.H., Wu, P.: Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of the ACM International Conference on Multimedia, pp. 157–166 (2014)
42. Yianilos, P.N.: Data structures and algorithms for nearest neighbor search in general metric spaces. In: Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 311–321 (1993)
43. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. Adv. Neural Inf. Process. Syst. **27**, 487–495 (2014)