Enrichment of News Show Videos with Multimodal Semi-Automatic Analysis

Daniel Stein[1], Evlampios Apostolidis[2], Vasileios Mezaris[2], Nicolas de Abreu Pereira[3], Jennifer Müller[3], Mathilde Sahuguet[4], Benoit Huet[4], and Ivo Lašek[5]

Fraunhofer Institute IAIS, Sankt Augustin, Germany[1] Information Technologies Institute CERTH, Thermi-Thessaloniki, Greece[2] Rundfunk Berlin-Brandenburg, Potsdam, Germany[3] Eurecom, Sophia Antipolis, France[4] Czech Technical University in Prague, and University of Economics, Prague, Czech Republic[5]

Abstract: **Enriching linear videos by offering continuative and related information via, e.g., audiostreams, webpages, as well as other videos, is typically hampered by its demand for massive editorial work. While there exist several (semi-)automatic methods that analyse audio/video content, one needs to decide which method offers appropriate information for an intended use-case scenario. In this paper, we present the news show scenario as defined within the LinkedTV project, and derive its necessities based on expected user archetypes. We then proceed to review the technology options for video analysis that we have access to, and describe which training material we opted for to feed our algorithms. Finally, we offer preliminary quality feedback results and give an outlook on the next steps within the project.**

Keywords: Speaker Recognition, Video Segmentation, Concept Detection, News show scenario, LinkedTV

1 Introduction

Many recent surveys show an ever growing increase in average video consumption, but also a general trend to simultaneous usage of internet and TV: for example, cross-media usage of at least once per month has risen to more than 59% among Americans [5]. A newer study [14] even reports that 86% of mobile internet users utilize their mobile device while watching TV. This trend results in considerable interest in interactive and enriched video experience, which is typically hampered by its demand for massive editorial work.

A huge variety of techniques, both new and established, exists that analyse video content (semi-)automatically. Ideally, the processed material will offer a rich and pervasive source of information to be used for automatic and semi-automatic interlinking purposes. However, the information produced by video analysis techniques is as heterogeneous as is their individual approach and the expected complexity, which is why careful planning is crucial, based on the demands of an actual use-case scenario. This paper introduces a news broadcast scenario for video enrichment as envisioned in the EU-funded project "Television linked to the Web" (LinkedTV),[1] and analyses the needs for practical usage. These needs form the basis for the technology that we employ for video analysis, and for the databases that we use to feed the training algorithms. We present our decisions on automatic speech recognition, keyword extraction, shot/scene segmentation, concept detection, the detection and tracking of moving and static objects, as well as unsupervised face clustering. Finally, we offer preliminary results based on human feedback.

This paper is structured as follows: we present the envisioned news show scenario of LinkedTV (Section 2), describe the analysis techniques and training material that are currently used (Section 3), provide manual examination of first experimental results (Section 4), and finally elaborate on future directions that will be pursued (Section 5).

1.1 Related Work

We will continue to review three recent projects with a related overall focus as LinkedTV. For the more detailed analysis methods as listed in Section 3, we will give appropriate citations along with their descriptions later in the paper.

inEvent The Project "Accessing Dynamic Networked Multimedia Events" (inEvent)[2] works on video material analysis and search-ability, using A/V processing techniques enriched with semantics, and recommendations based on social network information. The project's main targets are meetings, video-conferences and lectures, which are more restricted in a sense that they only include a limited set of persons and domains within one video.

TOSCA-MP "Task-oriented Search and Content Annotation for Media Production" (TOSCA-MP),[3] aims at content annotation and search tools, with its main target being professionals in the networked media production as well as archives, i.e., it is not directly aimed for non-professional end-users. The media scope is broader than LinkedTV, since TOSCA-MP also allows for radio pod-casts and written text from websites as seed content.

AXES The scope of the project "Access to Audiovisual Archives" (AXES)[4] is even broader, looking for possible linking information in scripts, audio tracks, wikis or blogs. A main focus is cross-modal detection of various entities such as people or places, i.e., drawing knowledge from several sources at once to improve the accuracy. The aimed content is audiovisual digital libraries rather than television broadcast.

2 News Broadcast Scenario

The audio quality and the visual presentation within videos found in the web, as well as their domains, are very heterogeneous. To cover many aspects of automatic video analysis, we have identified several possible scenarios for interlinkable videos within the LinkedTV project, e.g., (a) news show, (b) cultural heritage, and (c) visual arts [12]. In this paper, we focus on the specific demands of the news show scenario. The scenario uses German news broadcast as seed videos, provided by Public Service Broad-

[1] www.linkedtv.eu

[2] www.inevent-project.eu

[3] www.tosca-mp.eu

[4] www.axes-project.eu

caster Rundfunk Berlin-Brandenburg (RBB).[5] The main news show is broadcast several times each day, with a focus on local news for Berlin and Brandenburg area. For legal and quality reasons, the scenario is subject to many restrictions as it only allows for editorially controlled, high quality linking. For the same quality reason only links selected from a restricted whitelist are allowed. This whitelist contains, for example, videos produced by the Consortium of public service broadcasting institutions of the Federal Republic of Germany (ARD) and a limited number of approved third party providers.

The audio quality of the seed content can generally be considered to be clean, with little use of jingles or background music. Interviews of the local population may have a minor to thick accent, while the eight different moderators have a very clear and trained pronounciation. The main challenge for visual analysis is the multitude of possible topics in news shows. Technically, the individual elements will be rather clear: contextual segments (shots or stories) are usually separated by visual inserts and the appearance of the anchorperson, and there are only few fast camera movements.

2.1 Scenario Archetypes

LinkedTV envisions a service that offers enriched videos which are interlinked with the web, and targets a broader audience. For the sake of a convincing scenario, however, we have sketched three archetypal users of the LinkedTV news service and their motivations to use it:

Ralph comes home from working on a building site in Potsdam, and starts watching "rbb AKTUELL". The first spots are mainly about politics and about Berlin. Ralph is not particularly interested, neither in politics nor in Berlin as he lives in a small town in Brandenburg. After a while there is the first really interesting news for Ralph: a spot about the restoration of a church at a nearby lake; as a carpenter, Ralph is always interested in the restoration of old buildings. Therefore, he watches the main news spot carefully and views an extra video and several still images about the church before and after its restoration. Finally, the service also offers links to a map and the website of the church which was set up to document the restoration for donators and anyone else who would be interested. Ralph saves these links to his smartphone so he can visit the place on the weekend.

Nina's baby has fallen asleep after feeding, so she finds some time for casually watching TV, to be informed while doing some housework. Browsing the programme she sees that yesterday's enhanced "rbb AKTUELL" evening edition is available and starts the programme. Nina watches the intro with the headlines while starting her housework session with ironing some shirts. Watching a news spot about Berlin's Green Party leader who withdrew from his office yesterday, Nina is kind of frustrated as she voted for him and feels her vote is now "used" by someone she might not have voted for. She would like to hear what other politicians and people who voted for him think about his decision to resign. She watches a selection of video statements of politicians and voters and bookmarks a link to an online dossier about the man and his political carrier which she can browse later on her tablet. Eventually, the baby awakes so Nina pauses the application so she can continue later.

[5] www.rbb-online.de

Socially active retiree **Peter** watches the same news show with different personal interest and preferences. One of the spots is about a fire at famous Café Keese in Berlin. Peter is shocked. He used to go there every once in a while, but that was years ago. As he hasn't been there for years, he wonders how the place may have changed over this time. In the news spot, smoke and fire engines was almost all one could see, so he watches some older videos about the history of the famous location where men would call women on their table phones – hard to believe nowadays, he thinks, now that everyone carries around mobile phones! After checking the clips on the LinkedTV service, he returns to the main news show and watches the next spot on a new Internet portal about rehabilitation centres in Berlin and Brandenburg. He knows an increasing number of people who needed such facilities. He follows a link to a map of Brandenburg showing the locations of these centres and bookmarks the linked portal website to check some more information later. At the end of the show, he takes an interested look at the weather forecast, hoping that tomorrow would be as nice as today so he could go out again to bask in the sun.

3 Technical Background

Now that we have defined the motivation and needs of the archetypes above, we need access to a very heterogeneous set of information to be (semi-)automatically derived from the A/V content. We will proceed to list the technology employed to address these requirements. This section is divided into four sub-parts, being automatic speech recognition, temporal segmentation, spatiotemporal segmentation, person recognition, and other meta-information.

All this material is joined in a single xml file for each video, and can be visualized and edited by the annotation tool EXMARaLDA [9]. Note that for the time being, EXMARaLDA does not support spatial annotation; we plan to extend the tool at a later stage of the project. Also note that the EXMARaLDA format is only used internally and has been chosen because of its simplicity rather than a broad, standardized international usage (like in, e.g., MPEG-7).

3.1 Automatic Speech Recognition

Whenever there are no subtitles available, an automatic speech recognizer is needed in order to employ keyword extraction as well as named entity recognition. If subtitles are given, forced alignment techniques to match the timestamps to the video on a word level could be useful, because the utterance timestamp provided by the subtitles might be to imprecise and coarse-granular for our needs. In both cases, we need a strong German acoustic model. We employ a state-of-the-art speech recognition system as described in [10]. For training of the acoustic model, we employ 82,799 sentences from transcribed video files. In accordance with the news show scenario, they are taken from the domain of both broadcast news and political talk shows. The audio is sampled at 16 kHz and can be considered to be of clean quality. Parts of the talk shows are omitted when, e.g., many speakers talk simultaneously or when music is played in the background. The language model consists of the transcriptions of these audio files, plus additional in-domain data taken from online newspapers and RSS feeds. In total, the material consists of 11,670,856 sentences and 187,042,225 running words. Of these, the individual subtopics were used to train trigrams with modi-

fied Kneser-Ney discounting, and then interpolated and optimized for perplexity on a with-held 1% proportion of the corpus.

For Dutch, as foreseen in later parts of the project, the SHOUT speech recognition toolkit, as described in [6] will be used.

3.2 Temporal Segmentation

Larger videos should be temporally segmented based on their content. In our scenario, this would enable to skip parts of the video and directly jump to the weather forecast, for example. Also, we need to provide reasonable timestamp limits for hyperlinks, since we do not want further information on Berlin's Green Party to still be active once the clip about the rehabilitation center starts.

Currently, we segment the video into *shots* (i.e., fine-granular temporal segments which correspond to a sequence of consecutive frames captured without interruption by a single camera) and *scenes* (higher-level temporal segments composed by groups of shots, covering either a single event or several related events taking place in parallel).

Video shot segmentation is based on an approach proposed in [13]. The employed technique can detect both abrupt and gradual transitions between consecutive shots. The detection of gradual transitions is beneficial in cases where video production effects, such as fade in/out, dissolve etc., are used for the transition between successive shots of the video, which is a common approach, e.g., at the production of documentary videos. However, in certain use cases, like e.g., in news show videos where transition effects are rarely used between shots, it may be advantageous for minimizing both computational complexity and the rate of false positives to consider only the detected abrupt transitions. Specifically, this technique exploits image features such as color coherence, Macbeth color histogram and luminance center of gravity, in order to form an appropriate feature vector for each frame. Then, given a pair of selected successive or non-successive frames, the distances between their feature vectors are computed, forming distance vectors, which are then evaluated with the help of one or more SVM classifiers. In order to further improve the results, we augmented the above technique with a baseline approach to flash detection. Using the latter we minimize the number of incorrectly detected shot boundaries due to camera flash effects.

Video scene segmentation groups shot segments into sets which correspond to individual events of the video. The employed method was proposed in [11]. It introduces two extensions of the Scene Transition Graph (STG) algorithm [15]; the first one aims at reducing the computational cost of shot grouping by considering shot linking transitivity and the fact that scenes are by definition convex sets of shots, while the second one builds on the former to construct a probabilistic framework towards combination of multiple STGs. The latter allows for combining STGs built by examining different forms of information extracted from the video (i.e., low-level audio or visual features, and high-level visual concepts or audio events), while at the same time alleviating the need for manual STG parameter selection.

3.3 Spatiotemporal Segmentation

Objects of interest should be detected and tracked so that they can be clicked on for further information. Due to the broad target domains it cannot be guaranteed that established databases contain enough instances for local entities, which is why we need strong clustering and re-detection techniques so that an editor only needs to label them once and can automatically find other instances within the video itself, or within a larger set of surrounding videos.

For the purpose of spatiotemporal segmentation of the video stream, we differentiate between *moving* and *static* object detection. Spatiotemporal segmentation of a video shot into differently moving objects is performed as in [2]. This unsupervised method uses motion and color information directly extracted from the MPEG-2 compressed stream. The bilinear motion model is used to model the motion of the camera (equivalently, the perceived motion of static background) and, wherever necessary, the motion of the identified moving objects. Then, an iterative rejection scheme and temporal consistency constraints are employed for detecting differently moving objects, accounting for the fact that motion vectors extracted from the compressed stream may not accurately represent the true object motion.

For detecting occurrences of static objects of interest in consecutive or non-consecutive video frames, like the church or the café in the scenario, a semi-automatic approach based on object re-detection will be adopted. The human editor will manually specify the object of interest by marking a bounding box on one frame of the video. Then, the additional instances of the same object in subsequent frames will be automatically detected via object matching. Matching between image regions will be performed based on SURF descriptors [1] and some geometric restrictions defined by the RANSAC algorithm [2]. A baseline OpenCV implementation will initially be adopted for this. For each pair of images, feature vectors will be extracted using the SURF algorithm and will be compared. False matches will be filtered out using a symmetrical matching scheme between the pair of images, and the remaining outliers will be discarder by applying some geometric constraints calculated from the RANSAC method.

3.4 Person Detection

Persons seen or heard within a video are arguably the most important information for a viewer. They should be identified via their face and/or their voice, i.e., one can rely both on face detection as well as speaker recognition. For local content where some of the persons might be unknown with regard to the training material, there is also demand for unsupervised person clustering similar to objects of interest above, so that, e.g., the former leader of Berlin's Green Party needs only to be labelled once by a human editor.

3.4.1 Face Analysis

Face analysis is performed on keyframes extracted from the video (i.e., by temporal subsampling), and employs the face.com API.[6] The process can be divided into three components: face detection, face clustering and face recognition. Face detection is used as a prior tool to perform the other tasks, which both stem from the calculation of the similarity between detected faces. Our implementation is based on an iterative training/recognition approach [3] to make accurate groupings: the training process is initialized with the first detected face in the video. For each subsequent picture, the detected faces are matched against the

[6] http://developers.face.com

initial face. If the recognition confidence level is higher than a threshold (80% performed well in our experiments), both faces are associated with the same id, otherwise a new face id is created. After every assignment, the corresponding face model is retrained before performing recognition on the next candidate face. We output our results in a xml file that provides the coordinate of the faces detected in each picture (center, length and width of the bounded-box), together with the id of the face (id of the cluster). In order to guaranty the highest accuracy possible, we rely on a human annotator to label the face clusters with the appropriate person name.

3.4.2 Speaker Recognition

For speaker identification (SID), we follow the approach of [7], i.e., we make use of Gaussian Mixture Models (GMMs) using spectral energies over mel-filters, cepstral coefficients and delta cepstra of range 2. An overall universal background model (UBM) is merged from gender-dependent UBMs and forms the basis for the adaptation of person-dependent SID models. To train and assess the quality of the speaker identification, we listed German politicians as a possible requirement within our scenario description in Section 2.1. Thus, we downloaded a collection of speeches from 253 German politicians, taken from the archive of the German parliament.[7] In total, this consists of 2581 files with 324 hours of training material. To make training of the models feasible, we use 2 minutes per file to adapt the UBM.

3.5 Meta-Information

For keywords needed to tag the videos, they can either be extracted from textual sources, or derived from video-based concept detectors. These tags can then be used to recommend similar videos like, e.g., of churches in the local area.

3.5.1 Concept Detection

A baseline concept detection approach is adopted from [4]. Initially, 64-dimension SURF descriptors are extracted from video keyframes by performing dense sampling. These descriptors are then used by a Random Forest implementation in order to construct a Bag-of-Words representation (including 1024 elements) for each one of the extracted keyframes. Following the representation of keyframes by histograms of words, a set of linear SVMs is used for training and classification purposes, and the responses of the SVM classifiers for different keyframes of the same shot are appropriately combined. The final output of the classification for a shot is a value in the range [0, 1], which denotes the Degree of Confidence (DoC) with which the shot is related to the corresponding concept. Based on these classifier responses, a shot is described by a 323-element model vector, whose elements correspond to the detection results for the 323 concepts defined in the TRECVID 2011 SIN task.[8] These 323 concepts were selected among the 346 concepts originally defined in TRECVID 2011, after discarding a few that are either too generic (e.g., "Eukaryotic Organism") or irrelevant to the current data being considered in LinkedTV (cf. [12]).

[7] http://webtv.bundestag.de/iptv/player/macros/bttv/index.html

[8] www-nlpir.nist.gov/projects/tv2011/

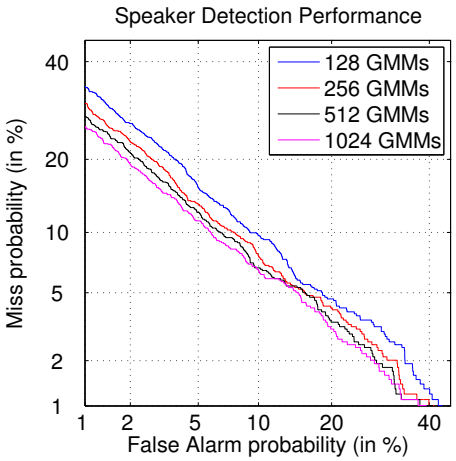

Figure 1: Speaker identifaction for German politians: DET Curve for different mixture sizes of the GMM, on a withheld test corpus of 994 audio files from the German parliament.

Moreover, in order to improve the detection accuracy we used some relations between concepts as they were determined in TRECVID 2011. When a concept implies another concept (e.g., the concept "Man" implies the concept "Person") then the confidence level of the second concept is reinforced with the help of an empirically set factor α. On the contrary, when a concept excludes another concept (e.g., the concept "Daytime Outdoor" excludes the concept "Nightime") and if the confidence score of the first concept is higher than the second, the first one is enhanced and the second one is penalized accordingly.

3.5.2 Keyword Extraction

The objective of keyword extraction or glossary extraction is to identify and organize words and phrases from documents into sets of glossary-items or keywords. For this particular scenario, we have access to several sources of textual information about a particular video. These include subtitles, manual annotations of videos, and finally the transcripts obtained from the ASR. Since LinkedTV is a multilingual project, we decided to refrain from using linguistically based, i.e., language dependent techniques here but employ a statistical approach based on text frequency – inverse document frequency (TF-IDF) [8] weights of words extracted from videos.

4 Experiments

In this section, we present the result of the manual evaluation of a first analysis of "rbb AKTUELL" videos.

The ASR system produces reasonable results for the news anchorman and for reports with predefined text. In interview situations, the performance drops significantly. Further problems include named entities of local interest, and heavy distortion when locals speak with a thick dialect. We manually analysed 4 sessions of 10:24 minutes total (1162 words). The percentage of erroneous words were at 9% and 11% for the anchorman and voice-over parts, respectively. In the interview phase, the error score rose to 33%, and even worse to 66% for persons with a local dialect.

To evaluate the quality of the SID, a distinct set of 994 audio files has been used to evaluate the quality of the mod-

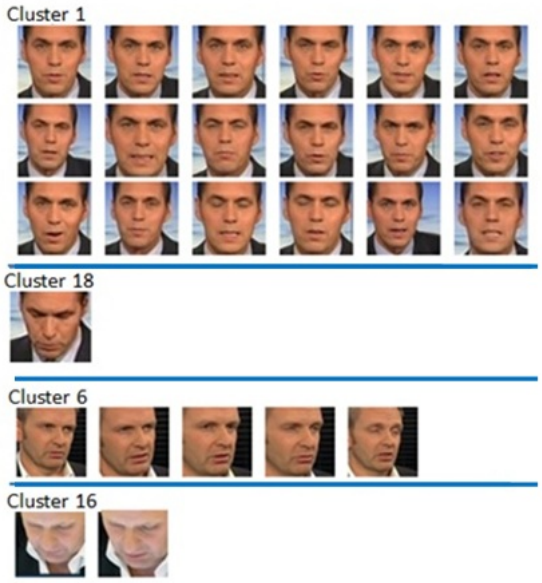

Figure 2: Clusters 1, 6, 16 and 18 of the face clustering result. Clusters 1 and 18 contain the anchorman and no noise face: two groups are made depending on the angle of his head. We have the same effect for clusters 6 and 16.

els, containing 253 different speakers. A GMM with 128 mixtures has an Equal Error Rate (EER) of 9.86, whereas using 1024 mixtures improves the EER to 8.06. See Figure 1 for Detection Error Trade-Off (DET) curves.

We performed face clustering on 200 keyframes extracted from the seed video at regular intervals. Results give a total of 54 clusters, most of which containing a single face (people that appear once). All 54 clusters are pure (i.e., contain only 1 person's face), and 2 ids (persons) appear in more than 1 cluster (as seen in Figure 2) due to significant viewing angle difference. Such clusters can easily be processed by an annotator. To make this process easier, we consider to let aside the clusters that contain only one face image, on the assumption that a person that appears only once is not of primary importance for the video.

In preliminary experiments on shot segmentation, the algorithm performed remarkably well. The effect from reporters' flashlights has been significantly restricted and the detection accuracy based on human defined ground-truth data was over 90%. A small number of false positives and false negatives was caused due to rapid camera zooming operations and shaky or fast camera movements. In a second iteration with conservative segmentation, most of these issues could be addressed.

Indicative results of spatiotemporal segmentation on these videos, following their temporal decomposition to shots, are shown in Figure 3. In this figure, the red rectangles demarcate automatically detected moving objects, which are typically central to the meaning of the corresponding video shot and could potentially be used for linking to other video, or multimedia in general, resources. Currently, the algorithm detects properly over 80% of the presented moving objects. However, an unwanted effect of the automatic processing is the false recognition of name banners which slide in quite frequently during interviews, which indeed is a moving object but does not yield additional information.

Figure 3: Spatiotemporal Segmentation on video samples from news show "RBB Aktuell"

Manually evaluating the top-10 most relevant concepts according to the classifiers' degrees of confidence revealed that the concept detectors often succeed in providing useful results; yet there is significant room for improvement. See Figure 4 for two examples, the left one with good results, and the right one with more problematic main concepts detected.

5 Conclusion

In this paper, we presented the news show scenario as pursued by the LinkedTV consortium. We have access to state-of-the-art techniques that can analyse the video content in order to derive the needed information, and reported reasonable preliminary results.

The main challenge will be to interweave the single results into refined high-level information. For example, the person detection can gain information from automatic speech recognition, speaker recognition and face recognition. Also, in order to find reasonable story segments in a larger video, one can draw knowledge both from speech segments, topic classification, and video shot segments. As a final example, video similarity can be estimated with feature vectors carrying information from the concept detection, the keywords, the topic classification and the entities detected within the video.

We already collected all (semi-)automatically produced information bits into a single annotation file that can be viewed with a proper tool, and will use this in order to establish ground truth material on the scenario data in a next step. Then, we plan to validate and extend the methods to increase their accuracy. Finally, we hope to gain more insight from the multimodal feature combination, which includes combining different confidence scores from across

Figure 4: Top 10 TREC-Vid concepts detected for two example screenshots. Wrong concepts in the left one are "outdoor","table","computers", and "clearing", whereas the rest is correct. In the right one, only three concepts "face", "body part" and "adult" are correct.

the analysis results to emphasize or reject certain hypotheses, but also will introduce high-level features that can offer new interlinking enhancements for the viewers' experience.

Acknowledgements

This work has been funded by the European Community's Seventh Framework Programme (FP7-ICT) under grant agreement n° 287911 LinkedTV.

References

[1] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359.

[2] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.

[3] Liu, X. and Huet, B. (2010). Concept detector refinement using social videos. In *VLS-MCMR — International workshop on Very-large-scale multimedia corpus, mining and retrieval, October 29, 2010, Firenze, Italy*, Firenze, ITALY.

[4] Moumtzidou, A., Sidiropoulos, P., Vrochidis, S., Gkalelis, N., Nikolopoulos, S., Mezaris, V., Kompatsiaris, I., and Patras, I. (2011). ITI-CERTH participation to TRECVID 2011. In *TRECVID 2011 Workshop*, Gaithersburg, MD, USA.

[5] Nielsen (2009). Three screen report. Technical report, Nielsen Company.

[6] Ordelman, R. J. F., Heeren, W. F. L., de Jong, F. M. G., Huijbregts, M. A. H., and Hiemstra, D. (2009). Towards Affordable Disclosure of Spoken Heritage Archives. *Journal of Digital Information*, 10(6).

[7] Reynolds, D., Quatieri, T., and Dunn, R. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10:19–41.

[8] Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 232–241, New York, NY, USA. Springer-Verlag New York, Inc.

[9] Schmidt, T. and Wörner, K. (2009). EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19:4:565–582.

[10] Schneider, D., Schon, J., and Eickeler, S. (2008). Towards Large Scale Vocabulary Independent Spoken Term Detection: Advances in the Fraunhofer IAIS Audiomining System. In *Proc. SIGIR*, Singapore.

[11] Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., and Trancoso, I. (2011). Temporal video segmentation to scenes using high-level audiovisual features. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(8):1163 –1177.

[12] Stein, D., Apostolidis, E., Mezaris, V., de Abreu Pereira, N., and Müller, J. (2012). Semi-automatic video analysis for linking television to the web. In *Proc. FutureTV Workshop*, pages 1–8, Berlin, Germany.

[13] Tsamoura, E., Mezaris, V., and Kompatsiaris, I. (2008). Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 45 –48.

[14] Yahoo! and Nielsen (2010). Mobile shopping framework – the role of mobile devices in the shopping process. Technical report, Yahoo! and The Nielsen Company. `l.yimg.com/a/i/us/ayc/article/mobile_shopping_framework_white_paper.pdf`.

[15] Yeung, M., Yeo, B.-L., and Liu, B. (1998). Segmentation of video by clustering and graph analysis. *Comput. Vis. Image Underst.*, 71(1):94–109.