

# Semi-Automatic Video Analysis for Linking Television to the Web

Daniel Stein<sup>1</sup>, Evlampios Apostolidis<sup>2</sup>, Vasileios Mezaris<sup>2</sup>, Nicolas de Abreu  
Pereira<sup>3</sup>, and Jennifer Müller<sup>3</sup>

<sup>1</sup> Fraunhofer Institute IAIS, Schloss Birlinghoven,  
53754 Sankt Augustin, Germany, [daniel.stein@iais.fraunhofer.de](mailto:daniel.stein@iais.fraunhofer.de)

<sup>2</sup> Informatics and Telematics Institute, CERTH,  
57001 Thessaloniki, Greece, [apostolid@iti.gr](mailto:apostolid@iti.gr) [bmezaris@iti.gr](mailto:bmezaris@iti.gr)

<sup>3</sup> rbb – Rundfunk Berlin-Brandenburg, 14482 Potsdam, Germany  
[nicolas.deabrepereira@rbb-online.de](mailto:nicolas.deabrepereira@rbb-online.de), [jennifer.mueller@rbb-online.de](mailto:jennifer.mueller@rbb-online.de)

**Abstract.** Enriching linear videos by offering continuative and related information via audiostreams, webpages, or other videos is typically hampered by its demand for massive editorial work. Automatic analysis of audio/video content by various statistical means can greatly speed up this process or even work autonomously. In this paper, we present the current status of (semi-)automatic video analysis within the LinkedTV project, which will provide a rich source of data material to be used for automatic and semi-automatic interlinking purposes.

## 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 [Nielsen, 2009]. A newer study [Yahoo! and Nielsen, 2010] 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. This paper introduces several scenarios for video enrichment as envisioned in the EU-funded project “Television linked to the Web” (LinkedTV),<sup>4</sup> and presents its workflow for an automated video processing that facilitates editorial work. The analysed material then can form the basis for further semantic enrichment, and further provides a rich source of high-level data material to be used for automatic and semi-automatic interlinking purposes.

This paper is structured as follows: first, we present the envisioned scenarios within LinkedTV, describe the analysis techniques that are currently used, provide manual examination of first experimental results, and finally elaborate on future directions that will be pursued.

---

<sup>4</sup> [www.linkedtv.eu](http://www.linkedtv.eu)

## 2 LinkedTV Scenarios

The audio quality and the visual presentation within videos found in the web, as well as their domains are very heterogeneous. To cover many possible aspects of automatic video analysis, we have identified several possible scenarios for interlinkable videos within the LinkedTV project, which are depicted in three different settings elaborated below. In Section 4, we will present first experiments for Scenario 1, which is why we keep the description for the other scenarios shorter.

**Scenario 1: News Broadcast** The first scenario uses German news broadcast as seed videos, taken from the Public Service Broadcaster Rundfunk Berlin-Brandenburg (RBB).<sup>5</sup> The main news show is broadcast several times each day, with a focus on local news for Berlin and the Brandenburg area. The scenario is subject to many restrictions as it only allows for editorially controlled, high quality linking with little errors. For the same quality reason only links selected from a restricted white-list are allowed, for example only videos produced by the Consortium of public-law broadcasting institutions of the Federal Republic of Germany. The audio quality can generally be considered to be clean, with little use of jingles or background music. Local interviews of the population might have a minor to thick accent, while the eight different moderators have a very clear and trained pronunciation. 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 there are only few quick camera movements.

**Scenario 2: Cultural Heritage** This scenario deals with Dutch cultural videos as seed, taken from the Dutch show “Tussen Kunst & Kitsch”, which is similar to the “Antique Roadshow” as produced by the BBC. The show is centered around art objects of various types, which are described and evaluated in detail by an expert. The restrictions for this scenario are very low, allowing for heavy interlinking to, e.g., Wikipedia and open-source video databases. The main visual challenge will be to recognise individual artistic objects so as to find matching items in other sources.

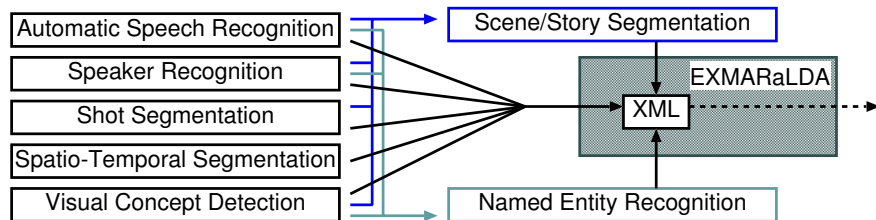
**Scenario 3: Visual Arts** At this stage in the LinkedTV project, the third scenario is yet to be sketched out more precisely. It will deal with visual art videos taken mainly from a francophone domain. Compared to the other scenarios, the seed videos will thus be seemingly arbitrary and offer little pre-domain knowledge.

---

<sup>5</sup> [www.rbb-online.de](http://www.rbb-online.de)



**Fig. 1.** Screenshots from the tentative scenario material: (a) taken from the German news show “rbb Aktuell” ([www.rbb-online.de](http://www.rbb-online.de)), (b) taken from the Dutch Art show “Tussen Kunst & Kitsch” ([tussenkunstenkitsch.avro.nl](http://tussenkunstenkitsch.avro.nl)), (c) taken from the Fire-TraSe project [Todoroff et al., 2010]



**Fig. 2.** Workflow of the data derived in LinkedTV

### 3 Technical Background

In this section, we describe the technologies with which we process the video material, and briefly describe the annotation tool as will be used for editorial correction of the automatically derived information.

In the current workflow, we derive stand-alone, automatic information from: automatic speech recognition (ASR), speaker identification (SID), shot segmentation, spatio-temporal video segmentation, and visual concept detection. Further, Named Entity Recognition (NER) and scene/story detection take the information from these information sources into account. All material is then joined in a single xml file for each video, and can be visualized and edited by the annotation tool EXMARaLDA [Schmidt and Wörner, 2009]. See Figure 2 for a graphical representation of the workflow.

**Automatic Audio Analysis** For training of the acoustic model, we employ 82,799 sentences from transcribed video files. They are taken from the domain of 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 modified Kneser-Ney discounting, and then interpolated and optimized for

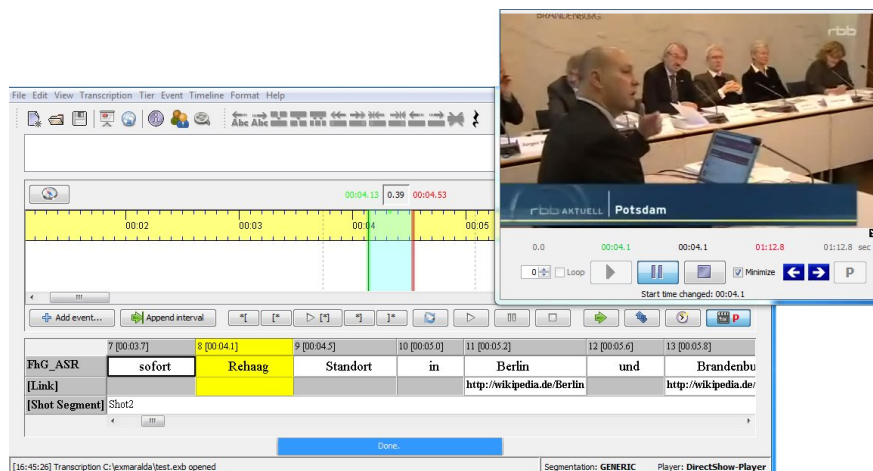
perplexity on a with-held 1% proportion of the corpus. The architecture of the system has been described in [Schneider et al., 2008].

**Temporal Video Segmentation** Video shot segmentation is based on an approach proposed in [Tsamoura et al., 2008]. The employed technique can detect both abrupt and gradual transitions; however, in certain use cases (depending on the content) 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 cameras flash effects.

Video scene segmentation draws input from shot segmentation and performs shot grouping into sets which correspond to individual scenes of the video. The employed method was proposed in [Sidiropoulos et al., 2011]. It introduces two extensions of the Scene Transition Graph (STG) algorithm; 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 multiple STG combination. The latter allows for combining STGs built by examining different forms of information extracted from the video (low level audio or visual features, visual concepts, audio events) while at the same time alleviating the need for manual STG parameter selection.

**Spatiotemporal Segmentation** Spatiotemporal segmentation of a video shot into differently moving objects is performed as in [Mezaris et al., 2004]. 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. Finally, both foreground and background spatiotemporal objects are identified.

**Concept Detection** A baseline concept detection approach is adopted from [Moumtzidou et al., 2011]. Initially, 64-dimension SURF descriptors are extracted from video keyframes by performing dense sampling. These descriptors are then



**Fig. 3.** Screenshot of the EXMARaLDA GUI, with automatically derived information extracted from an rbb video.

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. Random Forests are preferred instead of Nearest Neighbor Search, in order to reduce the associated computational time without compromising the detection performance. 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 346-element model vector, whose elements correspond to the detection results for the 346 concepts defined in the TRECVID 2011 SIN task.<sup>6</sup>

**Annotation** We make use of the “Extensible Markup Language for Discourse Annotation” (EXMARaLDA) toolkit [Schmidt and Wörner, 2009], which provides computer-supported transcription and annotation for spoken language corpora. EXMARaLDA supports many different OS and is written in Java. Currently, it is maintained by the Center of Speech Corpora, Hamburg, Germany. Here, the editor will have the possibility to modify automatic analysis results or add new annotations manually. This will be necessary especially with regard to contextualisation which still involves qualitative data analysis to some, if lesser, extent [de Abreu et al., 2006].

<sup>6</sup> [www-nlpir.nist.gov/projects/tv2011/](http://www-nlpir.nist.gov/projects/tv2011/)



Fig. 4. Spatiotemporal Segmentation on video samples from news show “RBB Aktuell”

## 4 Experiments

In this section, we present the result of the manual evaluation of a first analysis of videos from Scenario 1.



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.

In preliminary experiments on shot segmentation, zooming and shaky cameras were misinterpreted as the beginnings of new shots. Also, reporters’ flashlights confused the shot detection. In general, the algorithm detected 306 shots, whereas a human annotator found roughly one third, i.e., 103 shots. 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 4. 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, 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.

Manually evaluating the top-10 most relevant concepts according to the classifiers’ degrees of confidence, produced further room for improvement. As mentioned earlier, for the detection we have used 346 concepts from TRECVID, but evidently some of them should be eliminated for being either extremely specialized or extremely generic (“Eycariotic Organisms” is an example of an extremely generic concept), which renders them practically useless for the analysis of the given videos. See Table 1 for two examples. Future work on concept detection includes exploiting the relations between concepts (relations such as “man” being a specialization of “person”), in order to improve concept detection accuracy.

**Table 1.** Top 10 TREC-Vid concepts detected for two example screenshots from scenario 1, and their human evaluation

	Concept	estimation
	body part	good
	graphic	comprehensible
	charts	wrong
	reporters	good
	person	good
	face	good
	primate	unclear
	text	good
	news	good
	eukaryotic organism	correct
	event	unclear
	furniture	unclear
	clearing	comprehensible
	standing	wrong
	talking	good
	apartment complex	unclear
	anchorperson	wrong
	sofa	wrong
	apartments	good
	body parts	good

## 5 Conclusion

While still at an early stage within the project, the manual evaluation indicates that a first waystage of usable automatic analysis material could be achieved. We have identified several challenges which seem to be of minor to medium complexity. In the next phase, we will extend the analysis on 350 videos of a similar domain, and use the annotation tool to provide ground-truth on these aspects for quantifiable results. In another step, we will focus on an improved segmentation: since interlinking for a user is only relevant in a certain time intervall, we prepare to merge the shot boundaries and derive useful story segments based on simultaneous analysis of the audio and the video information, e.g., speaker segments correlated with background changes.

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

## References

- [de Abreu et al., 2006] de Abreu, N., Blanckenburg, C. v., Diemel, H., and Legewie, H. (2006). *Neue wissensbasierte Dienstleistungen im Wissenscoaching und in der Wissensstrukturierung*. TU-Verlag, Berlin, Germany.
- [Mezaris et al., 2004] Mezaris, V., Kompatsiaris, I., Boulgouris, N., and Strintzis, M. (2004). Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):606 – 621.
- [Moumtzidou et al., 2011] 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.
- [Nielsen, 2009] Nielsen (2009). Three screen report. Technical report, Nielsen Company.
- [Schmidt and Wörner, 2009] Schmidt, T. and Wörner, K. (2009). EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19:4:565–582.
- [Schneider et al., 2008] 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.
- [Sidiropoulos et al., 2011] 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.
- [Todoroff et al., 2010] Todoroff, T., Madhkour, R. B., Binon, D., Bose, R., and Paesmans, V. (2010). FireTraSe: Stereoscopic camera tracking and wireless wearable sensors system for interactive dance performances - Application to “Fire Experiences and Projections”. In Dutoit, T. and Macq, B., editors, *QPSR of the numediart research program*, volume 3, pages 9–24. numediart Research Program on Digital Art Technologies.
- [Tsamoura et al., 2008] 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.
- [Yahoo! and Nielsen, 2010] Yahoo! and Nielsen (2010). Mobile shopping framework – the role of mobile devices in the shopping process. Technical report, Yahoo! and The Nielsen Company.