From Raw Data to Semantically Enriched Hyperlinking: Recent Advances in the LinkedTV Analysis Workflow

Daniel Stein[1], Alp Öktem[1], Evlampios Apostolidis[2], Vasileios Mezaris[2],
José Luis Redondo García[3], Raphaël Troncy[3], Mathilde Sahuguet[3], Benoit Huet[3]

Fraunhofer Institute IAIS, Sankt Augustin, Germany[1] Information Technologies Institute CERTH, Thermi-Thessaloniki, Greece[2] Eurecom, Sophia Antipolis, France[3]

Abstract: **Enriching linear videos by offering continuative and related information via, e.g., audio streams, web pages, as well as other videos, is typically hampered by its demand for massive editorial work. While a large number of analysis techniques that extract knowledge automatically from video content exists, their produced raw data are typically not of interest to the end user. In this paper, we review our analysis efforts as defined within the LinkedTV project and present the recent advances in core technologies for automatic speech recognition and object-redetection. Furthermore, we introduce our approach for an automatically generated localized person identification database. Finally, the processing of the raw data into a linked resource available in a web compliant format is described.**

Keywords: Automatic Speech Recognition, Object Redetection, Person Identification, NERD Ontology

1 Introduction

Enriching videos (semi-)automatically with hyperlinks for a sophisticated viewing experience requires analysis techniques on many multi-modal levels. In [13], we presented the overall architecture decision for video analysis in the "Television linked to the Web" (LinkedTV)[1] project.

This paper focuses on the issues that we identified as most pressing (and there were quite a few): Local Berlin interviews featured a lot of interviews with the local residents, whose spontaneous speech produced only moderate automatic speech recognition (ASR) results. Speaker identification, while working properly on German parliament speeches, proved to be of little help since we had no localized database of Berlin speakers, a challenge that is shared with face recognition techniques. Object re-detection, for semi-automatically recognizing and tracking important objects in a show such as a local church or a painting, was too slow to be realistically employed in the architecture. Finally, the actual process of hyperlinking was left open in the last paper. In this follow-up paper, we present the new methods and the advances made, and explain our efforts in transforming raw data to semantically enriched and linked content.

This paper is organized as follows. After a brief description of the LinkedTV project (Section 2), we re-visit the ASR performance, which clearly showed deficiencies in spontaneous speech [13]. It has now been adopted to the seed content domain using a huge amount of new training material and a gradient-free optimization of the free decoding parameters (Section 3). Then, we present a stronger and faster solution for object re-detection (Section 4). Next, by interweaving several technologies such as face detection, video OCR and speaker identification, we can come up with a strong localized database for person identification (Section 5). Last, we elaborate on the actual hyperlinking stage, where the raw data is further processed (Section 6). Finally, we give a conclusion in Section 7.

[1] http://www.linkedtv.eu

2 LinkedTV

The vision of LinkedTV is of a ubiquitously online cloud of Networked Audio-Visual Content decoupled from place, device or source. The aim is to provide an interactive multimedia service for non-professional end-users, with focus on television broadcast content as seed videos. The project work-flow can be described as follows: starting from the demands of the use case scenarios, coupled with a description of the targeted multimedia content, the videos are analyzed by various (semi-)automatic ways. The raw data obtained from the single approaches is gathered and further enriched in a second step, by assigning media fragment descriptions and interlinking these with other multimedia information, using knowledge acquired from, e.g., web mining. The enriched videos are then shown in a suitably tailored presentation engine which allows the end-user to interact with a formerly linear video, and a recommendation/personalization engine which further gives the possibility to customize this experience.

In [13] we focused on the first two steps in this workflow, namely use case scenario and intelligent video analysis. There, we identified Berlin local news shows as seed content for the *news* use case, and the show "Tussen Kunst en Kitsch"[2] (similar to the Antiques Roadshow of the BBC), shown by Dutch public broadcaster AVRO,[3] as seed content for the *documentary* use case. This paper elaborates on the intelligent video analysis and the linking step as well as their interaction with each other.

3 ASR on Spontaneous Speech

Spoken content is one of the main sources for information extraction on all our relevant seed data sets. In [13], we performed a manual ASR transcript evaluation which performed good on planned speech segments, but rather poor on spontaneous parts which were quite common in interview situations in the news show scenarios. We thus decided to extend our training material with new data and adopt the settings of our decoder.

Recently, we collected and manually transcribed a huge new training corpus of broadcast video material, with a volume of approx. 400 h and containing roughly 225 h of clean speech. The new corpus is segmented into utterances

[2] http://www.tussenkunstenkitsch.nl

[3] http://www.avro.nl

Table 1: WER results on the test corpora, for the SPSA iterations and their respective loss functions. Each optimization on a given loss function has been executed two times from scratch with 18 iterations to check for convergence.

parameter set	WER planned	WER spontaneous
baseline	27.0	52.5
larger training data	26.4	50.0
SPSA 1st run	24.6	45.7
SPSA 2nd run	24.5	45.6

with a mean duration of 10 seconds and is transcribed manually on word level. The recorded data covers a broad selection of news, interviews, talk shows and documentaries, both from television and radio content across several stations. Special care has been taken that the material contains large parts of spontaneous speech. As the effort for acquiring new training data is still ongoing, the final size of the corpus will eventually reach 900 h, making this one of the largest corpora of German TV and radio broadcast material known to us.

This new training material made a revisit of the free speech decoder parameters necessary, to guarantee optimality. In the literature, these parameters are often either set empirically using cross-validation on a test set, which is a rather tedious task, or the default values of toolkits are retained. Few publications analyze the parameter adaption with automatic methods; among them are [3], using gradient descent, [7], using large-margin iterative linear programming, or [5], using evolutional strategies. Since we aim at facilitating the optimization process by employing a fast approach and therefore enable this step for a wide range of applications, we employ Simultaneous Perturbation Stochastic Approximation (SPSA) [12] for optimizing the free decoding parameters and show in [14] that it leads to stable and fast results.

The algorithm works as follows. For a tuple of free parameters in each iteration, SPSA perturbates the given values simultaneously, both adding and subtracting a random perturbation vector for a total of two new tuples. The gradient at the current iteration is estimated by the difference of the performance (here measured as word error rate, WER) between these two new tuples, and a new tuple is then computed by adapting the old tuple towards the gradient using a steadily decreasing step function. We refer to [14] for further implementation details.

For developing and optimizing the free parameters, we use a corpus from German broadcast shows, which contains a mix of planned (i.e., read news) and spontaneous (i.e., interviews) speech, for a total of 2,348 utterances (33,744 words).

For evaluation, we test the decoding performance on the news show content, separated into a planned set (1:08h, 787 utterances) and a spontaneous set (0:44h, 596 utterances). The results are listed in Figure 1. Here, it can be seen that while the performance for planned speech improved by 2.5% absolute (9.3% relative) in terms of WER, spontaneous speech segments now have a WER of almost 7% lower (13.3% relative) than the original baseline, which is quite a nice advance in the ASR quality.

4 Fast Object Re-detection

Since the videos in the presentation engine shall contain interactive (i.e. clickable) objects of interest, we need to associate visual content with appropriate labels. These labels can be automatically generated at the object-class level via high-level concept detection (by detecting concepts such as “car”, “person”, “building”, etc.), where we follow the approach of [10] using a sub-set of the base detectors described there. Moreover, a semi-automatic instance-based annotation of the video can be performed via the re-detection of specific objects of interest selected by the video editor so that, e.g., instances of the same painting in the antique road-show can be identified and tracked throughout the movie, allowing the viewer to click on them for further information or related videos.

We detect instances of a manually pre-defined object of interest O in a video V by evaluating its similarity against the frames of this video, based on the extraction and matching of of SURF (Speeded UP Robust Features) descriptors [2]. The time performance of our method is a crucial requirement, since the object-based video annotation will be handled by the editor. A faster than real-time processing is achieved by combining two different strategies: (a) exploit the processing power of the modern Graphic Processing Units (GPUs) and (b) introduce a video-structure-based frame sampling strategy that aims to reduce the number of frames that have to be checked.

Regarding the first strategy, GPU undertakes the initial decompression of the video into frames, the extraction and description of the image’s features and the matching of the calculated descriptors for a pair of images. Specifically, for the detection and description of the salient parts of the image a GPU-based implementation of the SURF algorithm is used, while the following matching step is performed in a brute force manner (i.e. each extracted descriptor from the object O is matched against all the extracted descriptors from the i-th frame F_i) looking each time for the 2-best matches via a k-Nearest Neighbor search for k = 2. This means that, for each detected interest point of O, the algorithm searches for the two best matches in F_i that correspond to the two nearest neighbors N_1 and N_2.[4]

The next steps aim to filter out any erroneous matches and minimize the incorrect (mis-)detections. Since they have lower computational complexity, they are handled by the Central Processing Unit (CPU). After matching descriptors between a pair of images, erroneous matches are discarded by applying the following rule: keep an interest point in O and its corresponding best match in F_i iff:

$$\|DistN_1\|_1 / |DistN_2\|_1 \leq 0.8,$$

where $\|\|_1$ is the Manhattan distance between the interest point in O and each of the calculated nearest neighbors. Additional outliers are then filtered-out by estimating the homography between O and F_i using the RANSAC algorithm [4]. If a sufficient number of pairs of descriptors remains after this geometric validation step, then the object is said to be detected in F_i and an appropriate bounding box is calculated and stored (i.e. the coordinates of the upper-left corner (x, y) and its width and height) for this frame, while otherwise the algorithm stores a bounding box of the form [0 0 0 0]. When the processing of the video frames is completed, a final filtering step is applied on the overall

[4] These GPU-based processes are realized using code included in version 2.4.3. of the OpenCV library, `http://www.opencv.org`

detection results aiming to the minimization of false positives (i.e. erroneous detections) and false negatives (i.e. erroneous misses). The latter is based on a sliding window of 21 frames and a set of temporal rules that decide on the existence or absence of the object O in the middle frame of this window.

Regarding the second strategy towards faster than real-time processing, further degradation of the needed processing time is achieved by designing and applying an efficient sampling strategy, which reduces the number of frames that have to be matched against the object of interest. The algorithm utilizes the analysis results of the shot segmentation method of [15], which can be interpreted as a matrix S where its i-th row $S_{i,j}, j = 1,...5$ contains the information about the i-th shot of the video. Specifically, $S_{i,1}$ and $S_{i,2}$ are the shot boundaries, i.e. the indices of the starting and ending frames of the shot and $S_{i,3}$, $S_{i,4}$, $S_{i,5}$ are the indices of three representative key-frames of this shot. By using this data, the algorithm initially tries to match the object O with the 5 frames of the i-th shot that are identified in matrix S (i.e. $S_{i,j}, j = 1,...5$), and only if the matching is successful for at least one of these frames it proceeds with comparing O against all the frames of that shot. It then continues with the key-frames of the next shot, until all shots have been checked. Following this approach the algorithm analyses in full only the parts (i.e. the shots) of the video where the object appears (being visible in at least one of the key-frames of these shots) and quickly rejects all remaining parts by performing a small number of comparisons, thus leading to a remarkable acceleration of the overall procedure.

More details on our object re-detection approach can be found in [1].

Our experiments on the object re-detection technique, using objects and videos from the LinkedTV dataset, show that the algorithm achieves 99.9% Precision and 87.2% Recall scores, identifying successfully the object for a range of different scales and orientations and when it is partially visible or partially occluded (see for example Fig. 1), while the needed processing time using a modest modern PC (e.g. having an Intel i7 processor, 8GB RAM memory and a CUDA-enabled GPU) is about 10% of the video's actual duration, thus making the implemented technique an efficient tool for fast and accurate instance-based annotation of videos within the LinkedTV analysis pipeline.

5 Towards Localized Person Identification

In the LinkedTV scenarios, object re-detection is one of the most important techniques in the documentary scenario, while person identification is far more crucial for the news show scenario. In [13], we described the challenge of obtaining a reasonable person identification database for local context. To overcome this, we exploit the fact that for most news show, banner information is shown whenever a specific person is interviewed. Manually checking videos of one show over the course of two months, it seems reasonable to assume that (a) the banner is only shown when the person is speaking, and (b) mostly – but not always – only this single person is seen in these shots. We can thus use this information for speaker identification and face recognition (cf. Figure 2 for a graphical representation of this work flow).

Figure 1: Object of interest (top row) and in green bounding boxes the detected appearances of it, after zoom in/out (middle row) and occlusion-rotation (bottom row).

For the show "Brandenburg aktuell"[5], we downloaded 50 videos over the course of two month, with each of 30 minutes length. Each show contains on average around seven interviewed persons with their name contained in the banner. Since the banner will be always at a certain position, we employ a simple yet effective Optical Character Recognition (OCR) heuristic using tesseract [11]: we check each screen-shot made every half second and decide that a name is found whenever the Levenshtein distance over three consecutive screen-shots is below 2. On manually annotated 137 screen-shots, the character accuracy is at convenient 97.4%, which further improves to 98.4% when optimizing tesseract on the shows font, using a distinct training set of 120 screen-shots.

This was used as a first reasonable basis for a speaker identification (SID) database. To obtain the audio portions of a speaker in a news excerpt, the banner is time-aligned to the speaker clustering segment, and other segments which have been assigned to be the same speaker via un-supervised clustering are also aligned to the same data collection. 269 instances with banner information were detected. The length of the spoken parts for a speaker in one show varied between 6 and 112 seconds, for an average of 31 seconds. 32 speakers appeared in more than one video.

For SID, we follow the approach of [9], 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. For evaluation of the speaker identification, we took every speaker that appeared more than once (32 speakers total) and divided videos of the two months of video material into a 2:1 ratio for training and testing. See Figure 3 for a Detection error tradeoff (DET) curve. The Equal Error Rate (EER) at 10.0% is reasonably close to the performance of German parliament speaker recognition (at 8.5% EER) as presented in our previous paper [13], but with the benefit that it is now on in-domain speakers.

[5] http://www.rbb-online.de/brandenburgaktuell/

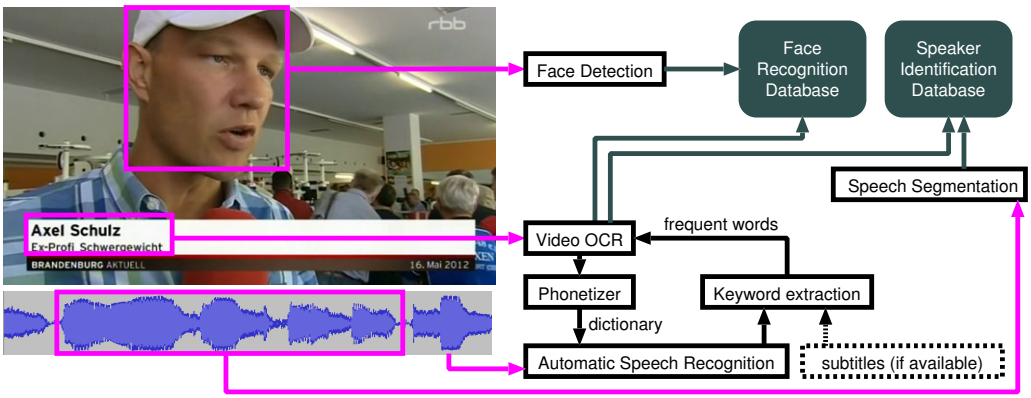

Figure 2: Workflow for a an automatically crawled person identification database, using news show banner information

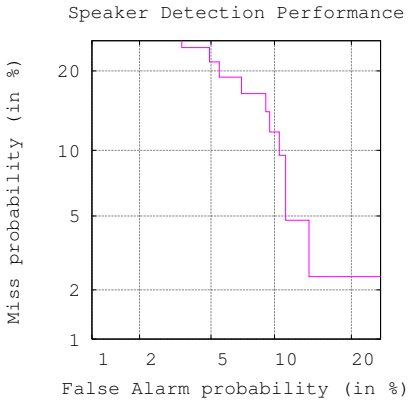

Figure 3: DET curve for the speaker identification experiment on RBB material.

In order to build a first database for face recognition, we applied face detection on the relevant screen-shots, using the widely used Viola-Jones detector [16], or more precisely its implementation in the OPENCV library as improved by Lienhart and Maydt [6]. Detection is combined with a skin color detector [8] for filtering out candidate regions that are not likely to be faces. Then, we link detected faces through shots using a spatio-temporal matching of faces: if two faces in adjacent frames are in a similar position, we assume we can match them. Interpolation of missing faces also relies on matching similar bounding boxes in close but none adjacent frames through a shot. This process enables to smooth the tracking results and to reject some false positive (when a track is too short, it is considered as a false alarm). See Figure 4 for the face detection results of one local politician that has been automatically harvested from the videos (he appeared in 11 different instances). These entries will serve as a database for face recognition in future work.

6 Hyperlinking

While in the previous sections we have focused on raw information extraction, this sections explains how the outcome from the visual and audio analysis performed over the video resources is transformed into a semantic graph representation, which enhances the way the information is exploited in a television scenario. The resultant Resource Description Framework (RDF) can be easier completed with other descriptions in external resources, better link-able with other content, and becomes available in a

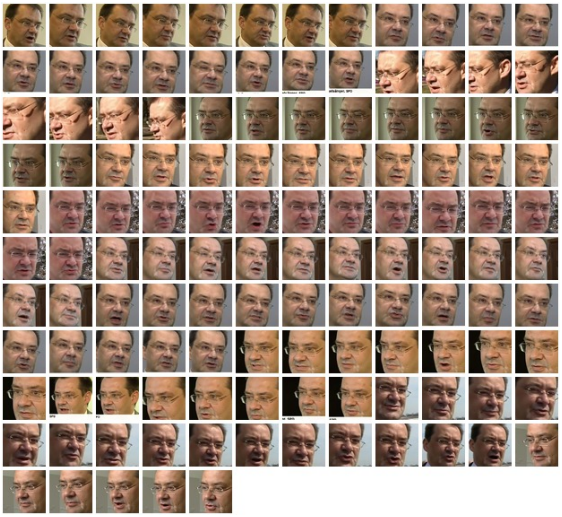

Figure 4: Crawled face shots from a local German politician, Jörg Vogelsänger.

Web compliant format that makes possible to bring hypermedia experience to the TV field.

RDF conversion In a first step, the aggregated information is converted into RDF and represented according to the LinkedTV Ontology[6]. The REST API service *tv2rdf*[7] performs this operation. The video content is structured in parts with different degrees of granularity, by using the Media Fragments URI 1.0 specification. Those instances of the *MediaFragment* class are the anchors where the entities will be attached in the following serialization step. The media fragment generation introduces a very important level of abstraction that opens many possibilities when annotating certain parts of the analyzed videos and makes possible to associate to fragments with other metadata with temporal references. The underlying model also relies on other established and well known ontologies like the The Open Annotation Core Data Model[8], the Ontology for Media Resources[9] or the NERD ontology. Table 2 shows some statistics about the number of MediaFragment's created for a 55 minutes chapter of the show *Tussen Kunst* in which five spatial object have been detected.

Below is the Turtle serialization of a spatial object detected in the same Tussen Kunst en Kitsch video, accord-

[6] http://semantics.eurecom.fr/linkedtv
[7] http://linkedtv.eurecom.fr/tv2rdf
[8] http://www.openannotation.org/spec/core
[9] http://www.w3.org/ns/ma-ont

Table 2: Number of MediaFragment's generated during the RDF serialization process of a Tussen Kunst en Kitsch episode.

Serialized Item	N MediaFragment's
Shots&Concepts	448
Subtitles	801
Bounding Boxes	4260
Spatial Objects	5

ing to the LinkedTV ontology. As every object can appear various times during the show, a different MediaFragment instance is created for each appearance. The temporal references are encoded using the NinSuna Ontology[10].

```
<http://data.linkedtv.eu/spatial_object/faedb8be-8de4
    -4e33-8d8c-26b35629785e>
     a       linkedtv:SpatialObject ;
     rdfs:label "CERTH_Object-5" .

<http://data.linkedtv.eu/media/e2899e7f-67c1-4a08
    -9146-5a205f6de457#t=1492.64,1504.88>
     a       nsa:TemporalFragment , ma:MediaFragment
          ;
     nsa:temporalEnd "1504.88"^^xsd:float ;
     nsa:temporalStart "1492.64"^^xsd:float ;
     nsa:temporalUnit "npt" ;
     ma:isFragmentOf <http://data.linkedtv.eu/media/
          e2899e7f-67c1-4a08-9146-5a205f6de457> .
```

At the same time every appearance is composed of a sequence of square bounding boxes that demarcate the object position, which are also represented as a set of MediaFragments of lower duration. The spatial references are directly encoded in the URL following the Media Fragments URI specification. The fact that one spatial MediaFragment belongs to the entire scope of a particular object is specified through the property MA:ISFRAGMENTOF.

Finally, broadcasters normally make available metadata related to their TV content, which is also included in the RDF graph during the serialization process. This data normally contains general information about the video such as: title, description, tags, channel, category, duration, language, creation date, publication date, view, comment, and subtitles. The service tv2rdf implements the serialization of TVAnytime[11] files into RDF by using the Programmes Ontology.[12]

Name Entity Extraction After the RDF graph is built, certain nodes are populated with extra anchors to the Link of Data Cloud. Named entity extraction processes are performed over the transcripts of the TV content that are available in the subtitle files from the providers or in the ASR results. The tv2rdf REST service launches this task by relying on the *NERD Client*, which is part of the NERD[13] framework. A multilingual entity extraction is performed over the video transcript and the output result is a collection of entities related to each video. Hence, the entities are classified using the core NERD Ontology v0.5[14] and serialized in JSON format, so they have to be translated by *tv2rdf* into a RDF representation and attached to the right MediaFragment.

During serialization, both Dublin Core[15] and LinkedTV properties are used in order to specify the entity label, confidence and relevance scores, name of the extractor used in the named entity recognition process, entity type and disambiguation URI (in this case, a resource in DBPedia). Below there is an example of the Turtle serialization for the entity Jan Sluijters spotted in the same episode of Tussen Kunst.

```
<http://data.linkedtv.eu/entity/9f5f6bc5-fa3a-4de1-
     b298-2ef364eab29e>
  a nerd:Person , linkedtv:Entity ;
  rdfs:label "Jan Sluijters" ;
  linkedtv:hasConfidence "0.5"^^xsd:float;
  linkedtv:hasRelevance  "0.5"^^xsd:float ;
  dc:identifier "77929" ;
  dc:source "semitags" ;
  dc:type "artist" ;
  owl:sameAs<dbpedia.org/resource/Jan_Sluyters>.
```

Table 3: Number of entities per type extracted from the Tussen Kunst en Kitsch video.

NERD type	Entities
Person	37
Location	46
Product	3
Organization	30
Thing	22

For having a better understanding of the number of entities extracted in the example video, the Table 3 presents some statistics about the extracted entities per NERD type.

Enrichment In a third step, the named entities already incorporated into the data graph are used for triggering processes to retrieve additional media content in the Web. The logic for accessing the external datasets where this information can be collected is implemented inside the LinkedTV REST service MediaCollector.[16] It is here where the original RDF graph is enriched with extra content that illustrates and completes what is shown in the seed video.

MediaCollector gets as input the label of the entities spotted by NERD over the transcript, and provides as result a list of media resources (photos and videos) grouped by source. For this research work the considered sources are selected from a white list defined by the content providers, due to the editorially controlled nature of the scenario. Those sources include mainly corporative Web Sites and some particular video channels in Youtube that have been previously checked by experts. When serializing the information, every item returned by MediaCollector is represented as a new MediaResource instance according to the Ontology for Media Resources. The entity used as input in the media discovery process is linked to the retrieved items through an OA:ANNOTATION instance, as proposed in the Open Annotation Ontology.

Data Exploitation Once the metadata about a particular content has been gathered, serialized into RDF, and interlinked with other resources in the Web, it is ready to be used in the subsequent consumption phases like the editorial review or data display. The creation of a MediaFragments hierarchy with different levels of granularity provides a very flexible model for (1) easily incorporate new data describing the media resource and (2) allowing different interpretations of the available information depending on the final user and the particular context.

For example, the spatial objects detected and named entities can be aligned for obtaining new insights about

[10] http://multimedialab.elis.ugent.be/organon/ontologies/ninsuna
[11] http://tech.ebu.ch/tvanytime
[12] http://purl.org/ontology/po
[13] http://nerd.eurecom.fr/
[14] http://nerd.eurecom.fr/ontology/nerd-v0.5.n3
[15] http://dublincore.org/documents/2012/06/14/dces
[16] http://linkedtv.eurecom.fr/api/mediacollector/

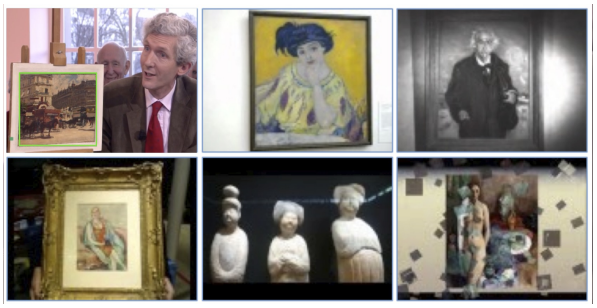

Figure 5: List of media items retrieved from MediaCollector service for the search term "Jan Sluijters".

what is happening in the video. The upper left image in Figure 5 illustrates a painting, detected by the object re-detection algorithm and highlighted with a green bounding box, that appears in the Tussen Kunst en Kitsch show, between the 1492nd and 1504th second. Looking for information attached to temporarily similar MediaFragments in the model, there is an entity about the artist "Jan Sluijters" that is mentioned from the second 1495 to 1502. So it is possible to conclude that this person is the author of the painting or at least is strongly related with it. Similar deductions can be done by relying in other items in the model like keywords and LSCOM concepts. The remaining images in Figure 5 correspond to some of the media items retrieved for the entity "Jan Sluijters". Most of them are about the relevant paintings created by this author.

Finally, as the resulting RDF graph is stored in an standard and Web compliant way, it can be used not only to be visualized in the LinkedTV platform but also for being referenced and consumed by other similar systems consuming television information.This way it is possible to implement solutions that bring innovative hyper-media experiences to the TV scenario.

7 Conclusion

In this paper, we presented recent improvements and strategies in the LinkedTV work-flow.

Generally speaking, the main challenge for harvesting semantically rich information from raw video input in sufficient quality is a matter of domain adaptation. We have shown ways to adopt the free decoder parameters to the new domain, requiring only a little amount of training data. Further, we presented improvements in the object re-detection algorithm which allows a fast and reliable detection and tracking of interesting objects. In order to obtain knowledge about the faces and voices of local people, we opted to crawl local news shows which usually contain banner information. We have shown that it is possible to build up a reasonable database fast, using well-established technology. Last, we showed how all this data is incorporated into the LinkedTV hyperlinking layer.

While there are many challenges up ahead, a first breakthrough from a collection of raw analysis data towards a semantically enriched linking has been established. As a next step, we focus on (1) multi-modal topic segmentation for link expiry estimation, and (2) multi-modal person identification, combining the knowledge from face recognition and speaker identification.

Acknowledgments This work has been funded by the European Community's Seventh Framework Programme (FP7-ICT) under grant agreement n° 287911 LinkedTV. LinkedTV would like to thank the AVRO for allowing us to re-use Tussen Kunst & Kitsch for our research.

References

[1] Apostolidis, E., Mezaris, V., and Kompatsiaris, I. (2013). Fast object re-detection and localization in video for spatio-temporal fragment creation. In *Proc. MMIX Workshop at IEEE Int. Conf. on Multimedia and Expo (ICME)*, San Jose, CA, USA.

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

[3] El Hannani, A. and Hain, T. (2010). Automatic optimization of speech decoder parameters. *Signal Processing Letters, IEEE*, 17(1):95–98.

[4] 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.

[5] Kacur, J. and Korosi, J. (2007). An accuracy optimization of a dialog asr system utilizing evolutional strategies. In *Proc. Image and Signal Processing and Analysis*, pages 180–184. IEEE.

[6] Lienhart, R. and Maydt, J. (2002). An extended set of Haar-like features for rapid object detection. In *Proc. Image Processing*, volume 1, pages I–900 – I–903 vol.1.

[7] Mak, B. and Ko, T. (2009). Automatic estimation of decoding parameters using large-margin iterative linear programming. In *Proc. Interspeech*, pages 1219–1222.

[8] Rahim, N. A. A., Kit, C. W., and See, J. (2006). Rgb-h-cbcr skin colour model for human face detection. In *MMU International Symposium on Information and Communications Technologies (M2USIC)*, Petaling Jaya, Malaysia.

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

[10] Sidiropoulos, P., Mezaris, V., and Kompatsiaris, I. (2013). Enhancing video concept detection with the use of tomographs. In *Proc. of the IEEE International Conference on Image Processing (ICIP)*, Melbourne, Australia.

[11] Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, pages 629–633, Washington, DC, USA. IEEE Computer Society.

[12] Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37:3.

[13] Stein, D., Apostolidis, E., Mezaris, V., de Abreu Pereira, N., Müller, J., Sahuguet, M., Huet, B., and Lašek, I. (2012). Enrichment of News Show Videos with Multimodal Semi-Automatic Analysis. In *Proc. NEM-Summit 2012*, pages 1–6, Istanbul, Turkey.

[14] Stein, D., Schwenninger, J., and Stadtschnitzer, M. (2013). Improved speed and quality for automatic speech recognition using simultaneous perturbation stochastic approximation. In *Proc. Interspeech*, pages 1–4, Lyon, France. to appear.

[15] 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 *Proc. Image Processing*, pages 45 –48.

[16] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proc. Computer Vision and Pattern Recognition, CVPR*, volume 1, pages I–511 – I–518 vol.1.