Semantic Annotation and Automated Extraction of Audio-Visual Staging Patterns in Large-Scale Empirical Film Studies

Henning Agt-Rickauer[1], Christian Hentschel[1], and Harald Sack[2]

[1] Hasso Plattner Institute for IT Systems Engineering,
University of Potsdam, Potsdam, Germany
{henning.agt-rickauer,christian.hentschel}@hpi.de
[2] FIZ Karlsruhe - Leibniz Institute for Information Infrastructure,
Karlsruhe Institute of Technology, Karlsruhe, Germany
harald.sack@fiz-karlsruhe.de

1 Introduction

The study of audio-visual rhetorics of affect scientifically analyses the impact of auditory and visual staging patterns on the perception of media productions as well as the conveyed emotions. In the *AdA*-project[3], together with film scientists, we aim to follow the hypothesis of TV reports drawing on audio-visual patterns in cinematographic productions to emotionally affect viewers by large-scale corpus analysis of TV reports, documentaries and genre-films of the topos "financial crisis". As it can be observed in the past and current media coverage of the world-wide financial crisis, TV reports often employ highly emotionalizing staging strategies in order to convey a certain message to the audience. This is also true for public broadcasting agencies who claim to adopt journalistic objectivity. This project therefore aims to make transparent this hard to grasp opinion-forming level of audio-visual reporting and follows the hypothesis of audio-visual staging patterns always aiming at coining emotional attitudes.

So far, localization and description of these patterns is currently limited to micro-studies due to the involved extremely high manual annotation effort [4]. We therefore pursue two main objectives: 1) creation of a standardized annotation vocabulary to be applied for semantic annotations and 2) semi-automatic classification of audio-visual patterns by training models on manually assembled ground truth annotation data. The annotation vocabulary for empirical film studies and semantic annotations of audio-visual material based on Linked Open Data principles enables the publication, reuse, retrieval, and visualization of results from film-analytical methods. Furthermore, automatic analysis of video streams allows to speed up the process of extracting audio-visual patterns.

This paper will focus on describing the semantic data management of the project and the developed vocabulary for fine-grained semantic video annotation. Furthermore, we will give a short outlook on how we aim to integrate machine learning into the process of automatically detecting audio-visual patterns.

[3] AdA-project — http://www.ada.cinepoetics.fu-berlin.de

2 Tool-Supported Empirical Film Studies

The systematic empirical study of audio-visual patterns in feature films, documentaries and TV reports requires a digitally supported methodology to produce consistent, open and reusable data. The project relies on tool-based video annotation and data management strictly following the Linked Open Data principles. A recent survey on available video annotation tools that can output RDF data can be found in [5]. Advene [1] was chosen for this project as annotation software because it meets best the needs of film scientists: It offers timeline view and segment based annotations of video sequences using multiple tracks which can be used to annotate various aspects under which a video is analyzed. We also collaborate with the author of Advene to develop project-specific extensions.

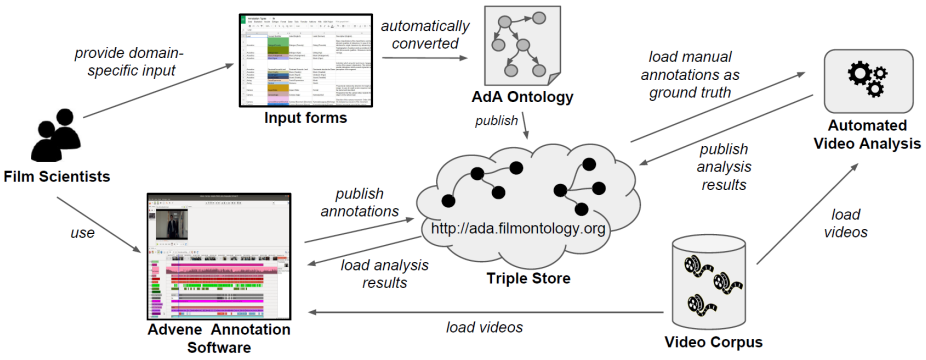

Fig. 1. AdA Project Setup — Semantic Annotation, Automated Video Analysis, and Linked Open Data Publication for Empirical Film Studies.

Fig. 1 shows the setup of the project from a data management perspective. At the beginning of the project a series of films was preselected and a corpus on the subject of the financial crisis was compiled. In the first phase, a part of the corpus is annotated manually using Advene based on film-analytical methods. To build a standardized annotation vocabulary from these methods, we have created spreadsheet-based forms that allow film scientists to provide domain-specific concepts, terms, and descriptions using a familiar software environment. We have developed an automated process to generate the project ontology and semantic metadata of the video corpus directly from the input data using the RDF mapping language and RML tools[4]. The ontology is imported into Advene and exposes the domain-specific vocabulary with unique URIs as annotation tracks in a timeline view so that semantic annotations conforming to the ontology can be exported. Annotations, metadata, and the ontology is published via the project's triple store[5]. The project is developing several services for automated

[4] `http://rml.io/index.html`
[5] `http://ada.filmontology.org/`

video analysis to reduce the need for elaborate manual annotations (analysis in the background or triggered from the Advene software, see Sect. 4).

3 Vocabulary for Fine-Grained Semantic Video Annotation

Movies are annotated on a specific topic using an annotation method (eMAEX[6]) developed by and especially for film scientists. eMAEX enables a precise description of the cinematographic image in its temporal development, which leads to several hundred annotations per scene. One goal of the project is to make these annotations accessible as Linked Open Data for exchange and comparison of analysis data. Therefore we have developed an ontology data model that uses the latest Web Annotation Vocabulary[7] to express annotations and Media Fragments URI[8] for timecode-based referencing of video material. The film scientists provide domain-specific knowledge for the development of a systematic vocabulary of film-analytical concepts and terms (the AdA Ontology), as existing ontologies such as LSCOM[9], COMM[10], MediaOnt[11], VidOnt [6] do not provide the level of detail required to describe all the low-level features of audio-visual content in the project according to scientific film-analytical standards.

Our vocabulary provides nine categories for annoation, called **Annotation Levels**: segmentation, language, image composition, camera, montage, acoustics, bodily expressivity, motifs, and other optional aspects. Each of the levels contains several sub-aspects, referred to as **Annotation Types**. These types correspond one-to-one with the tracks in a timeline view of the annotation software. For example, analysis of *image composition* includes the annotation of brightness, contrast, color accents, and visual patterns, and annotations at camera level include several aspects of *camera movement*, such as type of movement, speed, and direction. Each of these types has its own properties, the so-called **Annotation Values**. The ontology provides, if applicable, a set of predefined annotation values for each annotation type. For example, the type of camera movement can take *pan, tilt, zoom*, and other values. About 75% of the annotation types provide predefined values. The others contain free-text annotations, such as descriptions of scene settings or dialogue transcriptions.

The current version of the AdA ontology includes 9 annotation levels, 79 annotation types and 434 predefined annotation values. We provide an online version of the ontology under `http://ada.filmontology.org/` and a download possibility under the project's GitHub page[12].

[6] Electronically-based Media Analysis of EXpressive movements — `https://bit.ly/2K8i368`
[7] `https://www.w3.org/TR/annotation-vocab/`
[8] `https://www.w3.org/TR/media-frags/`
[9] `http://vocab.linkeddata.es/lscom/`
[10] `http://multimedia.semanticweb.org/COMM/`
[11] `https://www.w3.org/TR/mediaont-10/`
[12] `https://github.com/ProjectAdA/public`

4 Classification of Audio-Visual Patterns

In order to speed up the annotation process, we apply computer vision and machine learning techniques for generating annotation data on unseen material. As a first step, all videos in the corpus are automatically segmented into shots based on video cuts (hard-cuts, fades and dissolves [2]). Individual shots are represented by one or more key-frames, depending on the shot length. Key-frames are used for further analysis of the visual video content. Colorspace analysis helps to identify important aspects about the image composition such as the dominating color palette, diverging salient color accents as well the overall brightness of a shot. Optical flow analysis can be used to classify camera movement into static, zoom, tilt and pan. Other important aspects such as the video content are harder to grasp. Visual concepts depicted in a video segment such as *landscape*, *person* or *skyscraper* have been successfully classified using deep convolution neural networks trained on large amount of manually labeled image data [3]. Since for this project the amount of training data is limited, we use transfer learning approaches to fine-tune a pretrained neural network to our target domain [7].

All automatically derived annotations will be published as RDF based on the Media Fragments standard. In order to distinguish them from manually generated annotations, provenience information and confidence scores are added.

Acknowledgments. This work is partially supported by the Federal Ministry of Education and Research under grant number 01UG1632B.

References

1. Aubert, O., Prié, Y.: Advene: an open-source framework for integrating and visualising audiovisual metadata. In: Proceedings of the 15th ACM international conference on Multimedia. pp. 1005–1008. ACM (2007)
2. Hentschel, C., Hercher, J., Knuth, M., et al.: Open Up Cultural Heritage in Video Archives with Mediaglobe. In: Proceeding of the 12th International Conference on Innovative Internet Community Systems (I2CS 2012). vol. 204 (2012)
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. Advances In Neural Information Processing Systems pp. 1097–1105 (2012)
4. Scherer, T., Greifenstein, S., Kappelhoff, H.: Expressive movements in audiovisual media: Modulating affective experience. In: Müller, C., Cienki, A., Fricke, E., et al. (eds.) Body – Language – Communication. An international handbook on multimodality in human interaction, pp. 2081–2092. De Gruyter Mouton, Berlin, New York (2014)
5. Sikos, L.F.: Rdf-powered semantic video annotation tools with concept mapping to linked data for next-generation video indexing: a comprehensive review. Multimedia Tools and Applications **76**(12), 14437–14460 (2017)
6. Sikos, L.F.: Vidont: a core reference ontology for reasoning over video scenes. Journal of Information and Telecommunication **2**(2), 192–204 (2018)
7. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? Advances in Neural Information Processing Systems pp. 1–9 (2014)