Between Data Mining and Human Experience – Digital Approaches to Film, Television and Video Game Analysis

Matthias Grotkopp (m.grotkopp@fu-berlin.de), Freie Universität Berlin, Germany und Thomas Scherer (thomas.scherer@fu-berlin.de), Freie Universität Berlin, Germany und Jasper Stratil (jasper.stratil@fu-berlin.de), Freie Universität Berlin, Germany und Henning Agt-Rickauer (Henning.Agt-Rickauer@hpi.de), Hasso-Plattner-Intitute, University of Potsdam, Germany und Christian Hentschel (christian.hentschel@hpi.de), Hasso-Plattner-Intitute, University of Potsdam, Germany und Jan-Hendrik Bakels (Jan.Bakels@fu-berlin.de), Freie Universität Berlin, Germany

Within the field of digital humanities, studies of extensive corpora still focus, by and large, on a combination of quantitative approaches and the epistemology of distant reading. While such an approach can generate entire new sets of research questions and perspectives with regard to the macrostructures of certain media, formats and genres, the underlying principles of abstraction, accumulation and statistics fall short when it comes to questions of performativity, dynamics of perception or aesthetic experience.

This circumstance becomes apparent in the field of temporal arts and media, especially if the respective research is shaped by a theoretical framework that draws on aesthetics and phenomenological approaches. In these cases, statistical data based on discrete entities are often of limited value (how many A-minor chords are featured in a piece of music? How often does a dancer perform a certain move? What is the occurrence rate for a long shot followed by a detail shot in a specific movie?). The very advantage of distant reading—stepping out of the tangible context of a certain point in time or space within a given work of art in order to get a grasp on overarching principles of the work as a whole or even larger corpora—turns into a dead end. While a subject matter that is being referred to in terms of a semiotic, semantic, or syntactic paradigm can be divided into discrete entities with a fixed ‘value‘ or ‘meaning‘, the experiential quality of a certain detail within a phenomenological approach to aesthetics and performativity largely depends on the aesthetic composition as a whole. Accordingly, temporal arts and media—within the scope of this panel: film, television, web videos and video games—pose a challenge to the ways the isolation of features, the encoding of media ‘texts,‘ and the accumulation of data are being conducted within the current methodologies available in digital humanities.

This panel features film scholars as well as computational scientists who have jointly developed a digital approach to compositional patterns in audiovisual moving images within the research group ‘Audio-visual rhetorics of affect’ (http://www.ada.cinepoetics.fu-berlin.de/en/). While the first paper reflects upon possible intersections between phenomenological film theory and digital tools for video annotation and analysis, the second paper introduces against this theoretical backdrop and by means of a concrete analysis, a systematic framework for video annotation and data visualization with regard to compositional patterns in audiovisual media. The third paper addresses how this systematic framework for video annotation can be further standardized and automatized through semantic data structures and semi-automatic video analysis tools developed within the computational sciences. The fourth paper sketches out how this approach to semi-automatic video annotation and analysis can help achieve an empirical perspective on video game practice.

1. Analyzing Audiovisual Images: Digital Humanities AND/OR Phenomenology?

Matthias Grotkopp

This paper aims to open our panel by making a case for an approach to digital film analysis that is grounded in film theory. How does one reconcile the inherent complexity of the object of study—the multimodal composition and temporal gestalt of audiovisual images—with a seemingly necessary reduction of complexity for data management structures?

In practice, this question has mostly been avoided: digitized audio, images and audiovisual images tend to be treated as a wealth of easily accumulable metadata that almost by accident have data attached to them that also can be issued for human perception. These metadata naturally supply valuable information to media practitioners, media historians, and analysts of social, economic and cultural context. They present complex networks of “who did what, when and where?” which can be explored in myriad ways (cf. Acland/Hoyt 2016). But what happens if one is interested in questions like: What is the specific mode of aesthetic experience shaped in a film scene, and how does it relate to similar or contrasting scenes across a corpus of films? What is the temporal unfolding of metaphors and other meaning-making processes? What is the affective process of a scene and what is the summary emotional effect that a film aims to create? What is the poetic logic of a film or its desired rhetorical impact?

This is where phenomenology comes in, since these are questions that can hardly be answered by quantitative measurement and statistical analysis of single parameters alone, like shot duration (cf. the well-established Cinemetrics-tool), color values, speech or tagging image content (to name only those that are most commonly addressed). Instead—to borrow a phrase from Maurice Merleau-Ponty—the task is to show "how something takes on meaning—not by referring to already established and acquired ideas but by the temporal or spatial arrangement of elements" (Merleau-Ponty 1964 [1948]: 57). And it is about locating this emergence of meaning in embodied processes of reception where technologically animated audiovisual moving images are realized and reflected as reconfigurations of the parameters of perception and cognition.

But what are the requirements for analytical tools and data evaluation that are based on such an approach, rooted in film theory and qualitative research, as compared to purely quantitative data mining? Here lies not only the practical problem of retaining and managing the complexity of the object, but also the theoretical problem that a) not all of the relevant parameters for analyzing audiovisual images can be quantified in a similar manner, that b) the temporal and spatial arrangement of multiple features is, for all intents and purposes, the primary property, and that c) these arrangements of features in the audiovisual data have to be conceptualized as realized in an embodied act of viewing.

This paper will therefore argue for a method of combining fine-grained, time-code-based annotation following a standardized routine, with queries and visualizations that do not target these annotations as data-to-be-processed, but as patterns and dynamics of viewing-processes.

2. Researching and Annotating Audiovisual Patterns – Methodological Considerations

Jasper Stratil and Thomas Scherer

Digital video annotation is often concerned with represented objects within the image, especially when it comes to (semi-) automated analysis of moving images (cf. Qi et al., 2007). However, while the automatic detection of a car or a face may be useful for some purposes, it is not sufficient when analyzing the expressive qualities of audiovisual images: what feelings are evoked by the audiovisual compositions? How do processes of meaning making unfold?

Following up on the theoretical implications of the phenomenological approach presented in the first paper of this panel, this paper is building on the fundamental difference between still and moving images, addressing a gap in video annotation practices: while a majority of (semi-) automatic annotation routines focus on the single-frame extraction, film and media theory emphasizes audiovisual images as time-based media that unfold as an expressive movement (Kappelhoff 2018). These temporal, cross-modal patterns are harder to detect and annotate than static phenomena deduced from individual frames. A pitfall of (semi-)automatic analysis lies in a bias that prioritizes those dimensions that can be easily detected automatically, but this can lead to a reconstruction of an experience that is closer to computer vision than to human experience.

A harsh, sudden and aggressive movement creates an entirely different experience than a soft gliding movement—similar to the unfolding of a melody (Scherer et al., 2014), and this is highly relevant for the rhetorical dimension of audiovisual images. How, then, can this dimension of audiovisual images be annotated?

Annotating movements that can be grasped at the level of compositional patterns requires a specific vocabulary that can only be acquired by combining theoretical and methodological discourses from film and media studies with the affordances of a systematic and machine-readable vocabulary in order to describe various different audiovisual dynamics (see also the third paper in this panel and ada.filmontology.org). The paper presents first results concerning this challenge that were developed in a close cooperation between film studies and computational sciences. It addresses questions of segmentation, multimodal description, the balance between precision and time efficiency, as well as the development and application of a systematic vocabulary. These results will be demonstrated by reference to films concerning the global financial crisis (2007–).

Through an exemplary film analysis the paper addresses the question of how this micro-perspective, i.e. the detailed reconstruction of compositional patterns, can be related to the macro-perspective of entire films. How is it possible to 'read' these patterns within the vast amount of data necessary for the micro-analysis of moving images? How can the process of film viewing be reconstructed by identifying patterns in video annotations? And how can recurring patterns be identified and related to each other – within a film or even across a group of films?

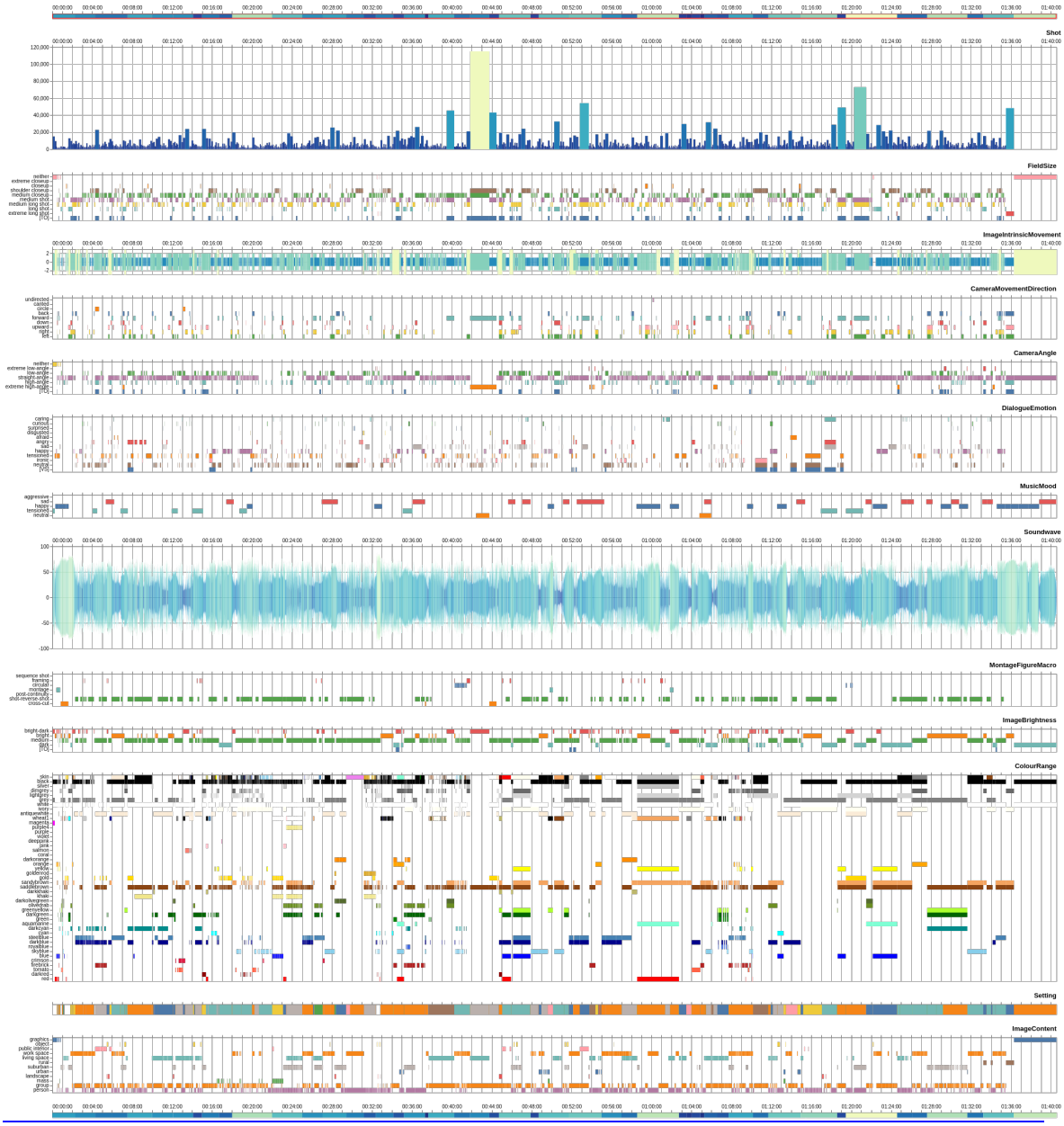

Figure 1. Exemplary view of a data set in ADVENE.

The paper will exemplify a use case of systematic digital video annotations and their visualization as a research tool for empirical film analysis that enables a scaling of detailed film-analytical approaches for a larger group of films, as well as the sharing of annotation data among scholars.

3. Standardization and Automation of Audiovisual Annotations

Henning Agt-Rickauer and Christian Hentschel

A holistic, scientific analysis of audiovisual patterns of staging in fictional and non-fictional video data requires dense localization and annotation of visual and auditory features within the video stream. Currently, the manual effort involved prevents analyses from going beyond micro-studies. On the one hand, this is partly because many annotation tools allow free-text annotations with arbitrary keywords and mapping the output of multimedia analysis tools to the desired target formats is costly (Kwasnicka & Jain 2018). On the other hand, film analysis, which aims to identify temporal and cross-modal patterns, requires the fine-grained annotation of a large number of film-analytical aspects involving several hours of work per minute of film. The approaches presented here, therefore, pursue two main objectives: 1) developing a standardized semantic annotation vocabulary for digital film analysis and 2) developing semi-automatic classification approaches of audiovisual patterns.

Manual video annotation performed by film scholars leads to highly valuable information that requires consistent management of video data, metadata, and annotation data in order to prevent them from being stored in isolated data silos and proprietary formats. We use Semantic Web technologies based on

Linked Open Data principles (Schmachtenberg et al., 2014) to publish, reuse, query, and visualize film-analytical data in order to share annotations with other film scholars and researchers from other disciplines. In this presentation, we will first show how semantic metadata of video files and movies help to create consistent video annotations for a large corpus. We will then show how concepts and terms from a film-analytical method are implemented as a semantic annotation vocabulary (Agt-Rickauer et al., 2018) and integrated with open source video annotation software allowing domain experts to author and publish unambiguous Linked Open Data. The presentation will also provide some insights into retrieval and analysis applications based on a large number of published annotations (http://ada.filmontology.org/).

Automatic analysis of video streams aims at improving the speed of localization and extraction of audiovisual features mainly through the application of approaches from computer vision and machine learning (Szeliski 2011). Shot boundaries and transition types can be automatically extracted by means of temporal video segmentation. Based on the extraction of content representing keyframes, the visual content of a shot can be further analyzed. Examples include extraction of dominant and salient colors, optical flow estimation for camera/object motion classification and visual concept detection for the identification of depicted objects and scenes. Finally, automatic analysis of the audio stream provides means for transcribing the spoken work into machine-readable text. The presentation will give a brief introduction into the main concepts behind these aforementioned approaches for semi-automatic annotation of audiovisual features with Linked Open Data.

4. What's Your Game? – A Digital Approach to Aesthetic Experience in Video Games

Jan-Hendrik Bakels

This paper suggests studying video games by means of digital film analysis in order to establish an empirical perspective on the act of playing, thereby addressing a crucial desideratum within game studies.

Though not contested in the past two decades as it has been in the past, the division into a narratological and a ludological perspective on video games still shapes the basic theoretical coordinates within the field of game studies. The narratologist approach considers video games to be situated within a larger framework of arts and media that are defined by their storytelling potentials, often focusing on the supposed closeness of games to the concepts of interactive and/or non-linear narration (Arseth 2012). In turn, the ludologist conception posits video games within the larger framework of game theory, stressing structural aspects such as rules (Frasca 2013). Most contributions to game studies take this opposition into consideration by referring to both aspects of video games, with different emphases (Juul 2005). This paper argues that, ultimately, both perspectives—at least from a phenomenological perspective—fall short to grasp the core of video game culture: the embodied experience of playing.

Drawing on phenomenology as well as theories on kinaesthesia and rhythm in performative arts and media, the paper attempts to develop an empirical-analytical approach to video gaming as an act of appropriation on behalf of the player. From this point-of-view, narratives are not to be grasped on the level of representation; instead, they derive from the complex interplay of a human being making an embodied experience and that human being's capability of narrating this experience. On the other hand, a game's rules and underlying structure merely shape the virtual potential of a game, not the way it is actually experienced while playing. In other words, if taken as theoretical vanishing points, both narratology and ludology fail to grasp video games at the level of actual playing: the first considers what is being 'taken away' from a game after the act, while the latter rather reflects upon the game in a virtual state before it is actualized by the act of playing.

Of course, kinaesthesia and embodiment have already been discussed in terms of video games before (Swalwell 2008). Nevertheless, at least one methodological problem contributes to preventing such a perspective from being elaborated in a systematic manner: in order to address aesthetics, embodiment, and the actualization of a game's potentialities at an analytical level, it is necessary to study actual acts of playing on the level of audiovisual images. But with most common video games lasting several hours, a comprehensive—not to mention: comparative—empirical study of games within the paradigm of media aesthetics seems to reach beyond the capacity of a single researcher or even research project. Against this backdrop, this paper will outline how the use of semi-automatic film-analytical tools developed in the past few years can provide the basis for a comprehensive and comparative study of specific acts of playing video games, thereby closing a crucial theoretical gap, as well as providing new impulses to the fields of narratology and ludology.

Appendix A

Bibliography

1. **Acland, C. R.; Hoyt, E.** (eds) (2016). *The Arclight Guidebook to Media History and the Digital Humanities* . Sussex: REFRAME Books.
2. **Agt-Rickauer, H., Hentschel, C. and Sack, H.** (2018). Semantic Annotation and Automated Extraction of Audio-Visual Staging Patterns in Large-Scale Empirical Film Studies. *SEMANTICS Posters & Demos 2018* .
3. **Arseth, E.** (2012). A Narrative Theory of Games. *Proceedings of the International Conference on the Foundations of Digital Games* , pp. 129–133.

4. **Frasca, G.** (2013). Simulation versus Narrative. Introduction to Ludology. In Wolf, M.J.P. and Perron, B. (eds), *The Video Game Theory Reader* . London & New York: Routledge, pp. 221–236.
5. **Juul, J.** (2005). *Half-Real: Video Games between Real Rules and Fictional Worlds* . Cambridge/MA: The MIT Press.
6. **Kappelhoff, H.** (2018). *Front Lines of Community: Hollywood Between War and Democracy* . Berlin/Boston: De Gruyter.
7. **Kwasnicka, H. and Jain, L.** (2018). Semantic Gap in Image and Video Analysis: An Introduction. In Halina K. and Lakhmi J. (eds), *Bridging the Semantic Gap in Image and Video Analysis* . Intelligent Systems Reference Library, pp 1–6.
8. **Merleau-Ponty, M.** (1964 [1948]). *The Cinema and the New Psychology. In Sense and Non-Sense* (Translated by Dreyfus, H. L. and Dreyfus, P. A.). Evanston/IL: Northwestern University Press, pp 48–59.
9. **Qi, G. J., Hua, X. S., Rui, Y., Tang, J., Mei, T. and Zhang, H. J.** (2007). Correlative multi-label video annotation. In *Proceedings of the 15th ACM international conference on Multimedia* . New York: ACM, pp. 17–26.
10. **Scherer, T., Greifenstein, S. and Kappelhoff, H.** (2014). Expressive movements in audiovisual media: Modulating affective experience. In Müller, C., Cienki, A. and Fricke E. (eds), *Body – Language – Communication. An international handbook on multimodality in human interaction (2)* . Berlin/Boston: De Gruyter Mouton, pp. 2081–2092.
11. **Schmachtenberg, M., Bizer, C. and Paulheim, H.** (2014). Adoption of the Linked Data Best Practices in Different Topical Domains. *International Semantic Web Conference (1)* .
12. **Sobchack, V.** (1992). *The Address of the Eye. A Phenomenology of Film Experience* . Princeton: Princeton University Press.
13. **Swalwell, M.** (2008). Movement and Kinaesthetic Responsiveness. A Neglected Pleasure. In Swalwell, M., Wilson, J. (eds) *The Pleasures of Computer Gaming: Essays on Cultural History, Theory and Aesthetics* . Jefferson/NC: McFarland, pp. 72–93.
14. **Szeliski, R.** (2011). *Computer Vision – Algorithms and Applications* . Texts in Computer Science. London: Springer.