DHQ: Digital Humanities Quarterly
2020
Volume 14 Number 4

Matching Computational Analysis and Human Experience: Performative Arts and the Digital Humanities

Jan-Hendrik Bakels <jan_dot_bakels_at_fu-berlin_dot_de>, Freie Universität Berlin
Matthias Grotkopp <mattjes_at_zedat_dot_fu-berlin_dot_de>, Freie Universität Berlin
Thomas Scherer <thomas_dot_scherer_at_fu-berlin_dot_de>, Freie Universität Berlin
Jasper Stratil <jasper_dot_stratil_at_fu-berlin_dot_de>, Freie Universität Berlin

Abstract

This article presents a framework that reconciles the requirements of computational methods with a qualitative, phenomenological approach to the analysis of audiovisual media. In its temporality and multimodality we treat audiovisual media as exemplary with regard to the wider field of performative arts and their analysis in digital humanities approaches.

First, we argue in favor of grounding digital methodology explicitly in scholarly, theoretical approaches to the human experience of performative arts and outline a qualitative approach to compositional patterns and dynamics of affect in audio-visual media. To demonstrate this approach, an exemplary scene analysis highlights the specifics of analyzing experiential qualities based on micro-level descriptions of compositional structures.

Eventually, the main body of the article spells out three central challenges with regard to this use of computational tools: 1.) recasting common film analytical vocabulary into a machine readable semantic ontology; 2.) setting up a systematic and applicable annotation routine that is based on the developed semantic ontology and allows for the interpersonal and consistent description of larger corpora; 3.) developing visualizations and query tools that enable the identification and tracing of compositional patterns within complex sets of annotation data.

The article concludes by demonstrating the benefits of visualized annotation data by taking up the exemplary analysis sketched out earlier and, ultimately, reflects upon the implications of the outlined AdA filmontology as a programmatic starting point to addressing intersubjective bases of experience in the wider field of digital humanities.

Research in the humanities concerned with aesthetics and performative arts draws upon a distinct tradition of what we might call “close reading:” Oscillating between formal description and hermeneutic interpretation, between shared myths, theories, codes and highly subjective lines of thought, the humanities’ analytical grasp on aesthetics and poetics is intimately tied to accounts of personal experience, rooted in the principles of exemplary study and subjective reading. 1

Over the past decades, developments within the field of computational science as well as the social and cultural shifts that are usually referred to with the broader term ‘digitalization‘ have affected core theory and methodology within the humanities. These developments provide alternatives and additions to familiar analytical approaches and give new weight to the kinds of statistical and numerical data collection that have been established with analogue means [Salt 1983]. New software-based methods have led to a whole range of research that aims at producing machine-readable data as well as processing this data within the logics of quantitative analysis, advanced statistics or visualization. With regard to audio-visual material, Cinemetrics [Tsivian 2009], the Digital Formalism project (e.g. Gruber et al. 2009) and especially Adelheid Heftberger [Heftberger 2018] as well as the film color projects conceived and led by Barbara Flückiger (e.g. Flückiger 2017), among others, have made groundbreaking contributions to this line of research. 2

At the same time, new possibilities for harvesting archives and metadata sets have propelled advances in historical research (e.g. Jacobs and Fyfe 2016, Verhoeven 2016). Large parts of the latter development have been described and 3

theorized convincingly under the term “distant reading” (see Moretti 2013). The methodology of distant reading has opened up new perspectives for research within the humanities, namely a shift from exemplary analysis to the comparative study of large corpora, while at the same time generating new modes of evidentness. For example, there are numerous approaches combining quantitative empiricism and humanities' epistemology [Schöch and Jannidis 2013] or producing new aesthetic modes of insight by means of big data visualization (e.g., the Cultural Analytics research by Manovich’s Software Studies Initiative [Manovich 2009] [Manovich and Douglas 2009] [Manovich 2016]) and other activities situated in the field of artistic and explorative research [Dawes 2004] [Ferguson 2015] [Ferguson 2017].

4 There are many possible research objectives that all share the feasibility of studying large corpora as a key requirement: for example the analysis of a very productive novelist’s body of work or the comparative study of all news reports on a given topic within a certain country and time frame. In particular the compilation of large sets of art works and media by means of social sciences and neuropsychology benefits from these kinds of insights from the humanities, may they be quantitative or – as we will try to demonstrate – qualitative. Therefore, the possibility of studying large corpora, which has been considered out of reach for single scholars or small groups of researchers who work within the framework of close and detailed hermeneutical studies, promises new perspectives for research within the humanities as well as interdisciplinary research.

5 But while opening up these perspectives, the methodology of distant reading – especially the turn to quantitative methods and often highly abstracting visualizations – has also created a major challenge with regard to advancing on this path in the field of art and media studies. These epistemological principles aim at expanding humanities' focus to larger corpora by an emphasis on the “distant” purposely transforming or even discarding the “reading.” It thus literally distances respective research in the field of the digital humanities from what we assume to be at the core of art and media consumption in the first place: the human experience of a given work of art or media. In this sense, “Distant reading is almost not reading at all” [Burdick et al. 2016, 39].[1] This mostly concerns studies of extensive corpora that by and large still focus on a combination of quantitative approaches and the epistemology of distant reading [Moretti 2013]. As a result, the methodology of distant reading has generated whole new sets of research questions and perspectives with regard to the macro structures and long term developments of certain media, formats and genres – but the underlying principles of abstraction[2], accumulation and statistics fall short when it comes to questions of performativity, dynamics of perception, or aesthetic experience.

6 This circumstance becomes apparent in the field of performative, time-based arts like music, theatre, dance or film – especially if the respective research is shaped by a theoretical framework that draws on phenomenological approaches to aesthetics and poetics. In these cases, statistical data based on discrete entities is often of limited value. It is well possible to count how many A-minor chords are featured in a given piece of music, how often a dancer performs a certain move, or which percentage of long shots are followed by a close-up in a film. But from a certain perspective within the study of performative arts, this information is epistemologically incomplete. With regard to aesthetic experience, it is only the specific tangible context – the harmonics and dynamics of that specific piece of music, the kinaesthetics of bodily expression in that certain dance routine or the interplay of music, cutting rhythm and acting within that particular film scene – which makes these features meaningful. 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) – shows its limits. Whereas a research object 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, these research objects pose a challenge to the ways the isolation of features, the encoding of media texts and the accumulation of data are currently being conducted within parts of the methodologies of digital humanities.

7 Against this backdrop, this paper addresses a simple question: Is it possible to use tools and methods developed within the fields of computational science and digital humanities in order to carry out qualitative analysis of aesthetic compositions? How can we productively study aesthetic experience while drawing upon the informational paradigm [Coppi 2002], i.e. the data driven operations of computational analysis?

In the following, we will present the methodological developments of the interdisciplinary project “Audio-visual rhetorics of affect”[3] (a cooperation of film studies and computational science; from here on referred to by its short title “AdA-project”) as an exemplary approach to the computational analysis of performative arts and media. Drawing on a qualitative approach to dynamics of affect in audio-visual media [Schmitt et al. 2014] [Scherer et al. 2014] [Kappelhoff and Bakels 2011] that combines a phenomenological understanding of film with the structural analysis of compositional patterns, this project aims at identifying rhetorical tropes within feature films, documentaries and TV reports on the global financial crisis (2007 and following). Addressing the questions posed above, this paper will focus less on specific findings (i.e. certain rhetorical tropes or a set thereof) but rather on the analytical framework that has been established in order to study a large corpus of audio-visual material with regard to aesthetic experience. This framework draws upon tools from the computational sciences, i.e. (semi-)automatic video analysis, semantic data structuring and machine learning.[4] 8

On the one hand, the AdA-project has to grasp its subject matter on the micro-level of audiovisual composition – due to the focus on audio-visual rhetorics and the dynamics of affect shaped by moving images. On the other hand, this micro-level has to become graspable within an epistemological framework similar to the approach of distant reading in order to identify recurring compositional patterns. 9

This path poses two obstacles: a) the micro-level description of compositional structures has to be carried out across a corpus of films and TV reports that is well beyond the scope of what film scholars are able to analyze by themselves within the timeline of a research project; b) an approach including the scalable analysis of aesthetics and expressivity has to generate paths towards modes of abstraction that are fundamentally different to what semiotic and statistical paradigms of distant reading currently have to offer, as well as to humanities analysis approaches based solely on natural language. Against this backdrop, this paper intends to treat the AdA-project’s methodology as an example for studying performative arts and media by means of computational analysis. By discussing the specific challenges the project has to address, the paper seeks to sketch out ways towards modes of analytical abstraction and data processing that remain close to the way artworks and media are experienced by human beings. It thus contributes to the more recent trend to integrate the close, the distant and the in-between under the terms of “scalable reading” or “scalable viewing” [Mueller 2012] [Pause and Walkowski 2019] [Fickers et al. 2018]. 10

In order to exemplify the path the project has taken so far with regard to the aforementioned obstacles, this paper will present a method from the field of film studies that – aiming at the affective experience of audio-visual sequences – qualifies their affective dimension by means of qualitative description and conceptually grasps the nexus of aesthetic experience of viewers and rhythmic-kinetic figurations of audiovisual images (see Sections 1.2 and 1.3). We will then proceed to address a number of challenges we encountered on our way of remodeling this method within a computational framework; this includes a detailed video-annotation routine, semantic data structures, the integration of tools that enable semi-automatic video analysis as well as computer-generated visualizations of compositional patterns. Therefore, the main focus of this article is on the recasting of the common knowledge of basic film analytical concepts into a consistent, machine-readable data model which provides the basis to address many challenges of digital film analytical methods (see Sections 2.1, 2.2 and 2.3). Finally, we will present how the analysis of the same example mentioned above unfolds if it is performed within the methodological framework of our computer-based approach (see Section 2.4) – and reflect upon the perspectives this kind of research offers with regard to matching computational analysis and human experience in the field of art and media studies (see Section 3). 11

1. A systematic approach to human experience: The eMAEX-Method and affective dynamics in film-viewing

Before we discuss this paper’s main question – how to analyze arts and media on the level of human experience within a digital humanities framework – it is important to address a crucial question on the level of film theory first: Is it after all possible to address aesthetic experience on a general level, regardless of individual dispositions and differences and without postulating universal mechanisms of perceiving and processing? This question concerns any kind of research dealing with human experience of temporal arts and media, not turning to methods of self-disclosure or physiological 12

measurements of actual empirical subjects (e.g. as applied within the social sciences and empirical psychology). In order to address it – in an exemplary way –, we are going to briefly sketch out a qualitative method developed within the field of film studies that aims at qualifying the affective experience of audio-visual segments based on film-analytical methods (see Kappelhoff and Bakels 2011, Kappelhoff 2018a, Scherer et al. 2014).

1.1 Aesthetic experience and intersubjectivity in film viewing – theoretical bases

13 As the programmatic remarks have indicated so far, we are strongly advocating for an approach to digital humanities that accentuates the theory driven and theory building aspect of humanities' research. The following chapter serves as a brief sketch of the film theoretical background that informs both our research questions and the direction of method and tool development in the collaboration with our colleagues from the computational sciences. Even though we hope that key elements of the approach – specifically the systematic annotation vocabulary and the data visualization – are employable by different theoretical schools, they have been developed in view of a framework that focuses on the premises of embodied perception and the expressivity of movement patterns.

14 From a film studies perspective, all methodological questions concerning film analysis in the context of digital humanities are preceded by the challenge to engage descriptively with a fleeting subject matter that solely exists within the time of its perception [Bellour 1975] [Bellour 2000], and more pinpointed to our theoretical perspective: its being-viewed by an embodied spectator. Some branches of film theory tend to avoid this problem by tying the spectator's emotions in film viewing to the cognitive apprehension of character and plot constellations [Tan 1996] [Grodal 1997] [Grodal 2009], largely leaving aside the media specific conditions of moving images and sound. While these theories have attracted a certain attention in the wake of a broader turn towards cognitive theory and neuropsychology over the past two decades, they remain in opposition to a theoretical strand that has been prominent within film studies since Hugo Münsterberg's early psychological accounts on the "photoplay" [Münsterberg 2002]. This line of research focuses on the dynamics of movement and rhythm as the central phenomena with regard to the aesthetic experience of moving images, especially with regard to questions of mood, feeling, affect or emotion. Within the early years of the 20th century, the concept of *expressive movement* began to serve as a crucial node at the intersection of art theory [Fiedler 1991], social philosophy [Simmel 1959] [Simmel 1993] and anthropology [Wundt 1880] [Wundt 1896] [Bühler 1933] [Plessner 1982]. While these theories referred to the concept mainly with regard to the human body's expressivity and its role in art and culture, Münsterberg and his successors in film theory – like Béla Balázs [Balázs 2010] or Sergej Eisenstein [Eisenstein 1991] – applied it to the kinetic and rhythmic patterns of the cinematic image.

15 In this tradition of thought, theories on movement and its expressive qualities have gained new attention following the *bodily turn* respectively *performative turn* within the humanities and social sciences that took place in the 1990s. In film theory, neo-phenomenological theories on *embodiment* and embodied perception once again turned the focus on movement and its crucial role at the intersection of expressivity and embodied perception (see Marks 2000, Barker 2009, Meunier 2019). Following Vivian Sobchack [Sobchack 1992], one of the most prominent voices within neo-phenomenological film theory, it is exactly this expressive quality of movement that provides the basis for an intersubjective understanding of experience in film viewing. According to Sobchack, the cinematic image presents itself as a situated seeing and hearing, a subjective perspective. In the act of film viewing, the spectator experiences the kinetic and haptic qualities of this situated viewing and hearing as an embodied being. The spectator's perception and the cinematic image – as an expression of perception, of seeing and hearing – are intertwined in a two-fold act of perception grounded in the principle of an embodied being's kinetic being-in-the-world:

> In a search for rules and principles governing cinematic expression, most of the descriptions and reflections of classical and contemporary film theory have not fully addressed the cinema as life expressing life, as experience expressing experience. Nor have they explored the mutual possession of this experience of perception and its expression by filmmaker, film, and spectator – all viewers viewing, engaged as participants in dynamically and directionally reversible acts that reflexively and reflectively constitute the perception of expression and the expression of perception. Indeed, it is this mutual capacity for and possession of experience through common structures of embodied existence, through similar modes of being-in-the-world, that provide the intersubjective

> basis of objective cinematic communication. [Sobchack 1992, 5]

Sobchack's notion of the intersubjectivity of kinetic and haptic dynamics of film resonates with thoughts on the commonalities of film and music Sergej Eisenstein has developed in the 1940s, suggesting specific compositional principles with regard to the arrangement of movement and rhythm while outlining his ideas of expressive movement and a cinematic aesthetics of effect [Eisenstein 1988]. Lately, these ideas on cinema's audio-visual musicality have again been picked up on in theories on audience engagement [Pearlman 2009] as well as with regard to dynamic affects in film-viewing [Kappelhoff 2004] [Kappelhoff 2018b] [Bakels 2017].

16 This conception of the audio-visual modulation of affects grounded in the temporal shaping of intensities and rhythms has also been argued for by appealing to the work of the developmental psychologist Daniel Stern [Stern 1985] [Stern 2010]. His theory of the vitality affects or vitality forms was introduced to the film theoretical discourse by Raymond Bellour [Bellour 2011]. Affects here are applied as self-contained temporal gestalts of movement, of rhythm, and of intensity, which are not linked to individual modalities of perception or forms of interaction. It is a matter of synaesthetic patterns, such as the creeping, the bulging, the explosion-like, or the fading, which can each occur as specific experiences, both in perception and in action as well as in feeling and thinking. These are derived from primordial forms of intersubjective, cross-modal interaction, like the affective reflection of an infant's facial expression in the voice of its mother. Cinema, according to Bellour, can also be thought of as a similar interaction, constantly translating perceptions into feelings, for it produces:

> in the variety of its components (the image and the modalities of the soundtrack incorporated in it) [...] the constant illusion of a sensory attunement between the elements of the world, just as it does between the bodies that are deployed in it. [Bellour 2011, 229]

17 Taken together, neo-phenomenological film theory's core assumption of the audio-visual image being intersubjectively experienced as a two-fold act of kinaesthetic perception on the one hand, and theories on expressive movement, rhythm and audio-visual composition on the other, offer a way to address the aesthetic experience of moving images from an analytical perspective. Theories on movement and rhythm in audio-visual images serve as a methodological starting point for reconstructing dynamics of swelling tensions, shifting kinetic forces and the temporal shapes that emerge within these dynamics – by means of systematically describing patterns of audio-visual composition and their experiential qualities. In turn, neo-phenomenological theories on the embodied perception of movement provide the theoretical basis for assuming these experiential qualities to be experienced regardless of individual dispositions and differences, i.e. addressing human experience on a more general level without claiming universality.

18 With regard to the question raised at the beginning of this section – whether it is at all possible to address aesthetic experiences in film viewing on a general level via means of film analysis – we can now give a more nuanced answer: Of course, we do not claim to be able to predict a specific human being's experience of a certain audio-visual sequence in detail and in total by means of film analysis. Obviously, we would never deny the notion of experience being rooted in specific cultural and historical contexts, as well as the conception that matters of feeling and understanding are highly subjective. Against the theoretical background we have sketched out here [Kappelhoff 2018b] [Müller and Kappelhoff 2018], we should be able to demonstrate to what extent embodied experiences of audio-visual sequences can be reconstructed – by referring analytically to the spatial and temporal dynamics of specific rhythmic-kinetic figurations of audio-visual composition. Or to put it differently: We all do make our own experience in watching a certain film, TV series or news report. But given the basic principles of rhythm, movement and embodied perception, these individual experiences should relate to commonly shared experiential dynamics that are reflected in one way or another within our individual accounts. The aim of the method outlined within the rest of this article is to grasp this commonly shared basis of film experience and to make it accessible for comparative research. But before we get to the question of how this goal is achieved via digital tools, we first like to present a methodology that was developed with regard to this theoretical framework and exemplify it with a short analysis.

1.2 eMAEX: qualifying the affective dynamics of audio-visual sequences by means of segmentation and description

19 The eMAEX [5] system aims at providing a methodological framework for analyzing and qualifying the expressive and affective qualities of audio-visual material – may it be a contemporary Hollywood feature film, arthouse cinema, a screwball comedy, a war documentary or a web video.[6] At the time of its initial development, it consisted of a systematic routine to segment, describe and qualify units of audio-visual composition as well as a web-based multimedia environment that enables researchers to combine analytical descriptions, film stills, data visualizations and the subject matter of analysis (i.e. video files of audio-visual source material) in order to display the results of film analytical studies. The computational approach to film analysis presented in Section 2 of this article is based on the analytical routine of the eMAEX framework and can be viewed as an advancement and adaptation of its routine in regard to the requirements of a semantically structured and machine-readable analytical vocabulary. Over the course of the following two subchapters of this section, we will exemplify the original eMAEX approach in order to a) unfold a vivid example of what we mean when referring to the aesthetic experience of moving images from a film studies perspective and b) thereby provide the necessary background for addressing this paper's central question: how a systematised and machine-readable analytical vocabulary in combination with digital tools for video annotation can help to expand such a perspective on embodied aesthetic experience with regard to the comparative analysis of large corpora.

20 Following the theoretical concepts of *expressive movement* and *embodiment* outlined above, eMAEX aims at the systematical description of spatio-temporal dynamics in moving images. Segmenting the subject matter at hand serves as the first step within the analytical routine and provides the basic temporal structure of the given subject matter. Following the intuitive elementary category of temporal segmentation within the everyday, journalistic and academic discourse, the object of study is first segmented into single scenes – conceptualized as an experiential unit.[7] Within the routine's vocabulary, this basic structure, the succession and affective interplay of scenes, is referred to as the object of research's temporal macro level.

21 While the segmentation into scenes is a methodological consensus in various film analytical approaches and often carried out intuitively, the second level of segmentation is intimately tied to the qualitative description of the analyzed material: Within the first step of analysis, an initial short description of the scene is produced; by focusing on five formal levels of audio-visual composition (namely choreography, camera, sound, gestures and facial expressions and image composition) rather than on narration and representation, this initial description provides a first insight into the compositional principles that shape the respective scene. The grouping of the many different aspects of audio-visual composition into these five levels was based on a pragmatic reasoning that balanced the need for some kind of standardization within projects with the limited possibilities of free text annotation. The elaboration of this standardization of description routines into a workable data modelling has been one of the main foci of our subsequent developments.

22 Based on this initial description, the scene as a whole is sub-segmented into consecutive *expressive movement units* (EMUs), i.e. the interplay of different levels of staging is described as *one temporal gestalt*, loosely comparable to a gesture of the film itself. Two of the five formal levels of composition mentioned above are chosen as dominant levels within each EMU. Usually changes with regard to dominant levels of composition provide strong hints with regard to the scene's segmentation into EMUs. Following this second step of segmentation, a qualitative description of the compositional dynamics shaping each EMU is produced separately, and the respective EMU is qualified with regard to its affective dynamics.

23 In a final, third step, the dynamic interplay of a scene's EMUs forming the compositional logic of the respective scene as a whole is also subject to a qualitative description of its affective dynamics. Following this last step, qualitative data on the respective subject matter of analysis has been gathered on three levels:

1. All scene's sub-segments of audio-visual composition (EMUs) are qualified with regard to their affective dynamics as cross-modal temporal gestalts. This in-depth analysis constitutes the micro-level of analysis.
2. The interplay of these sub-segments within each scene is also qualified with regard to its affective dynamics; this is referred to as the meso-level of analysis.
3. The qualifications with regard to each scene's affective dynamics are displayed as an affective course spanning the whole film; this course is referred to as the macro-level of analysis, revealing each research

object's basic dramaturgy of affect.

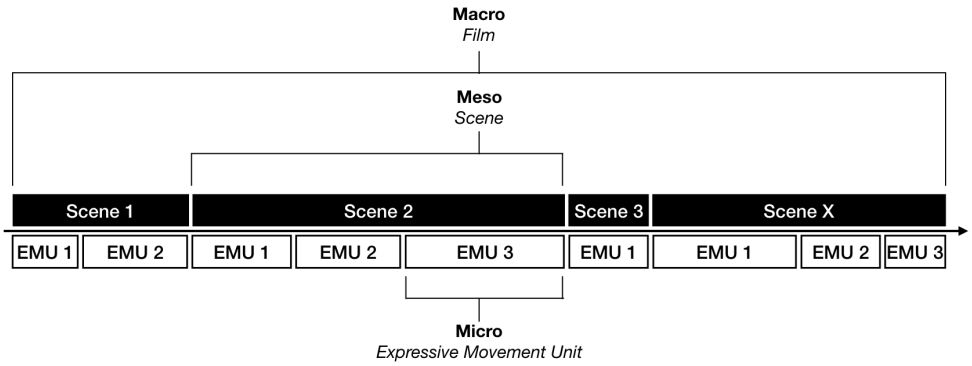

Figure 1. Macro, meso, and micro level of eMAEX.

The final subchapter of this section will unfold an exemplary scene analysis according to the eMAEX system in order to provide an example of how we conceive the aesthetic experience of audio-visual images – before the next section will address the question of how such a qualitative approach can be followed up upon via digital tools of video analysis.

1.3 Exemplary Analysis: A scene from *The Company Men*

In the following, we want to present the exemplary analysis of a scene from the AdA-project's corpus that is supposed to serve as a tangible example for the concept of cinematic expressive movement. The film *The Company Men* (John Wells, USA 2010) addresses the consequences of the financial crisis after 2007 in everyday life. More precisely, the film focuses on job cuts due to speculations on the stock market – by following several male protagonists and their struggle to find a new job in order to provide for their families.[8] 24

To give you a short overview of the exemplary scene, taken from the second half of the film: Bobby Walker – played by Ben Affleck –, who has been fired from his office job at a logistics company in Boston and lives with his family at his parent's house, is searching urgently for a job as marketeer. After getting a phone call with the invitation to a job interview and a short discussion with his wife, we see him preparing in a dark hotel room for the job interview with a company in Chicago: the protagonist awakes at dawn, carries out his fitness routine and irons his shirt. We then see him walking amid a crowd of people in business outfits towards a high corporate building. Inside the building, in a shiny lobby, Bobby walks towards the reception counter. After announcing himself he sits down and waits nervously. When the assistant of the awaited manager appears, Bobby learns that the appointment has been mixed up and that there won't be a job interview. The scene ends with Bobby standing disappointed in a busy crowd of passersby on the streets of Chicago. 25

In this plot summary – like in every film synopsis – the varying experiential qualities and dynamics of the scene have already been synthesized into attributions like "shiny," "nervously," "disappointed" and so on. The following analysis is aimed at grasping the expressive qualities in which those summaries – i.e. verbal accounts of experience – are rooted. In order to make substantiated claims about the scene's experiential quality, we will now break down the scene's compositional structure into EMUs, drawing on the second level of temporal segmentation of the eMAEX framework as outlined above. In doing so, we will briefly describe each EMU and its affective quality with regard to the temporal gestalt that is constituted through the interplay of its dominant formal levels – in this case acoustics (especially music), 26

image composition and camera.

First Expressive Movement Unit (TC: 01:11:41–01:12:36)

In the first EMU, a figuration of a sudden increase of vividness is created through the interplay of acting and acoustics, more specifically through harsh contrasts in the acting style – especially the bodily and facial expressivity of the main character played by Ben Affleck. At the beginning of the EMU, the fatigued protagonist walks up a driveway and enters his parents' home's kitchen. Laconic and with low body tension, he interacts with his family. When he's told that there was a call in his absence from a firm in Chicago offering him a job, the film immediately cuts to the protagonist packing his suitcase while discussing the job offer with his wife. His body tension rises and his gestures become more and more energetic and extensive as he talks animatedly to his wife. The intensity of the protagonist's movements are highlighted by the restraint of his counterpart: His wife leans on the side of the image, half in the dark with arms crossed and a smile, she only interrupts him with brief questions or affirming comments. The increasing vitality in acting ground the experience of an abruptly emerging and then continually rising energy.

Figure 2. Stills from The Company Men. EMU 1. TC: 01:11:41–01:12:36.

Second Expressive Movement Unit (TC: 01:12:36–01:13:26)

In a nonverbal sequence, we see Bobby preparing for the interview in his hotel room and walking through Chicago until he arrives at the company building. This process is staged as a swelling movement figuration: The music becomes louder and more upbeat and the image increasingly brightens. In the last part, the camera movements, Bobby's movement as well as those of passersby form a concerted choreography – an opening movement i.e. a cinematic gesture of widening. This choreography emphasizes the vertical axis and underlines the directedness of the whole sequence – from the hotel room to the busy streets of Chicago finally tilting up the tall corporate building. The interplay of music, movement and image composition form a swelling movement figuration that grounds an experience of vivid directedness.


4:15

Figure 3. Stills from The Company Men. EMU 2. TC: 01:12:36–01:13:26.

Third Expressive Movement Unit (TC: 01:13:26–01:14:37)

Through a repeated change between deep and flat image compositions as well as close and more distanced camera positions, Bobby's arrival in the waiting area of the company building is staged as an emerging tension which is supported through the hasty movements of Bobby fidgeting with a magazine. An assistant enters the image through a flight of stairs in the rear of the image and tells Bobby that his appointment will not take place. The lobby that the protagonist enters is staged as a space of suppressed vividness. This is achieved through the contrast of subtle, but constant movement of the isolated protagonists on the one hand and an image space dominated by geometric, rectangular shapes and clear lines and surfaces on the other. The camera approaches the conversation before cutting out abruptly to a medium long shot while the actors nearly freeze completely in their body movements and the dialogue stops. In this way, the interplay of acting, choreography and image composition grounds the experience of a vividness that is abruptly put to rest.


EXIT

Figure 4. Stills from The Company Men. EMU 3. TC: 01:13:26–01:14:37.

Fourth Expressive Movement Unit (TC: 01:14:37–01:14:56)

In a nonverbal succession of two shots, Bobby is standing again in the streets of Chicago. A stretched piano-and-strings arrangement sets in. The camera focuses on the protagonist, positioning him at the center of the image. His body's immobility is in harsh contrast to the many moving passersby that crowd the image in the fore- and background. Over two shots the camera moves towards the protagonist – thus narrowing the image space continually – and singles out his face that has lost its tension staring into the offscreen space. The contrast in movement exposes the protagonist and the swelling elegiac music emphasizes his facial expression. Together these articulatory strategies form a figuration of an intensifying deceleration and contraction that grounds an experience of slowly being isolated.

Figure 5. Stills from The Company Men. EMU 4. TC: 01:14:37–01:14:56.

The Scenes Dynamic Pattern as a Whole (TC 01:11:41–01:14:56)

The scene as a whole – drawing on the interplay of music, image composition and camera work – stages an affective parcours shaped by the experience of changing energetic states. Over the course of the whole scene, this affective course grounds an emerging image of disappointment. 31

After the initial sudden increase of vividness (the first EMU), a directed, swelling figuration (the second EMU) is abruptly put to rest (the third EMU). A figuration of intensifying deceleration and contraction (the fourth EMU) then closes the scene. 32

It is crucial that we are explicitly not talking about the emotions that the protagonist may feel, as if he was a person of flesh and blood – instead of being a mere composition of shapes, colors and sounds. Rather we claim that, in this scene, the unfolding of various audio-visual rhythms in the act of film-viewing [Kappelhoff 2018b] – i.e. the entanglement of technically animated movement images and embodied experiences – can be qualified and reconstructed as an image of disappointment; the affective course laid out over the description of the scene's four EMUs lends this image of disappointment its temporal and energetic shape – the abrupt rise of energy that comes with the increase of vividness; the directedness this joyful expectation can generate; the abrupt putting-to-rest that comes with disappointment; and finally the inwardness and isolation that shapes the aftermath of disappointment. All this is generated here on the level of the embodied experience of rhythms and gestures the audio-visual image performs. That way Bobby's disappointment is not simply represented, but rather made sensible to the viewer. 33

In this section we aimed at exemplifying the way in which performative arts and media directly address human capacities for embodied perception and sensation. While we hope this exemplary analysis helped in making this conceivable, succeeding in this case only means to highlight the actual problem: how the experiential dynamics described over the course of this exemplary analysis can be addressed within the informational paradigm of computational analysis. The following section will be dedicated to outlining challenges and perspectives in trying to achieve such a computational methodology. 34

2. Addressing human experience by means of computational analysis: The AdA-framework for video annotation 35

The exemplary analysis above sketched out a systematic, qualitative approach to describing and qualifying the 36

experiential qualities of audio-visual sequences that was developed in the field of film studies. Qualitative descriptions of expressive dynamics like this highlight the complexity as well as the variety of audio-visual composition: From lighting to camera movement to image composition to color grading to sound design to cutting rhythm to acting to choreographies and so on – the audio-visual image seems to encompass a sheer endless amount of formal levels. In addition, any compositional figuration of a multimodal audio-visual composition seems to realize a distinct way of making different of these countless levels interact. Furthermore, the foregrounding of certain dynamics within this interplay, the climaxes and accents that arise from it, are from our theoretical perspective tied to genuine phenomena of embodied perception. In short: Reconstructing these expressive dynamics of human experience in the way "traditional" scholarly work has operated – film-viewing and descriptions in natural language – seems to be resistant to any implementation as digital method or tool. However, we want to present a digital approach to film analysis in this section that is concerned exactly with these expressive dynamics and experiential qualities. The starting point in addressing these problems with regard to a data-driven computational approach to studies concerned with human experience turns out to be as simple as effortful: i.e. the task of producing extensive and detailed data on various different formal levels. The aim is to be able to retrace these expressive dynamics – as patterns – in a bottom-up perspective based on this data gathered on the micro level of audio-visual composition.

37 The turn to computational tools that shapes the methodological approach of our project is motivated by a research interest that highlights the practical limits of viewing and describing as a film analytical method.[9] Previous work in the many different exemplary case studies with the eMAEX methodology[10] has generated a tempting hypothesis: that the principles studied there yield generic systems of cinematic articulations on the micro-level of audio-visual composition and expressive dynamics. Taking the complexity and variety of these compositional principles into account, the concepts of repetition, variation and differentiation nevertheless open a path towards a typological study of expressive dynamics. Given the affective quality we attribute to these figurations of expressive movement, such a study aims at identifying a set of rhetorical tropes grounding audio-visual communication – i.e. at compositional gestures that mark the represented topics, constellations or concepts with affective qualities like tension, fear or euphoria, shaping the affective perspectives from which these topics emerge. To follow up on this hypothesis would require the comparative analysis of a large corpus of audio-visual material.

38 The AdA-project is dedicated to such an empirical study of a large corpus of audio-visual material – consisting of fictional feature films, documentaries and TV reports concerned with the global financial crisis (2007–) – with the aim of establishing a typological perspective on patterns of audio-visual expressivity. Given the discursive nature of crisis rhetorics – oscillating between the identification of a significant problem and the struggle for a solution – and the variety of audio-visual media, this corpus is considered suitable as an exemplary field of study. The project's methodological approach brings together film scholars and computational scientists; it encompasses

1. the development and definition of a film analytical ontology, i.e. a systematic analytical vocabulary that follows the requirements of machine-readable semantic data management, as well as film analytical key concepts that are widely used in film studies,
2. the annotation of extensive audio-visual material based on this vocabulary, combining tools for (semi-)automatic video analysis and manual annotations by expert annotators, as well as
3. evaluation and application of machine learning and/or search algorithms in order to identify recurring patterns of audio-visual composition, in combination with tools for visualizing and querying complex sets of annotation data for scholars (a step that is going to be evaluated in the last phase of our project).

39 The latter step as well as the use of (semi-)automatic tools integrated within the annotation software will follow the human-in-the-loop model, i.e. all data generated by computational analysis will be corrected by film scholars in order to further train the applied tools.

40 Out of the variety of tools for manual and semi-automatic video annotation[11], we have chosen the open-source video-annotation-software Advene, which was originally designed by Olivier Aubert, Yannick Prié and Pierre-Antoine Champin, to perform multi-author film analyses. Based on a cooperation with Olivier Aubert, Advene has been further adjusted and extended to meet the specific requirements of detailed scholarly film analysis – not only in regard to the

(manual) annotation process but also regarding interfaces for a film analytical ontology, video retrieval and the support of RDF (a standard model for data interchange on the Web; see Agt-Rickauer et al. 2018).

41 The latter functions as the exchange format for a machine-readable film-analytical ontology, i.e. the structured vocabulary and data modelling which is the basis for the annotation process. The framework is set up to make semantically stored data readable by machines and humans. While most annotations in the AdA-project are manually created, the amount of (semi-)automated annotations is still to be expanded further within the framework of this methodology. The aim is to develop a basis for the joint integration of manual and (semi-)automatic annotations and facilitate future work on machine learning for which we have laid the groundwork by providing a structured vocabulary and the interfaces in the annotation tool.

42 In the following, we will address the challenges that arise with developing the ontology, the implication of a consistent video-annotation routine, and the development of visualizations which offer scholars a way to 'read' into these extensive data sets. The section will be closed with a second look at the scene described at the end of the last section – now based on video annotation data and respective visualizations.

2.1 Challenge 1: Establishing a machine-readable vocabulary and data structure

43 The main methodological goal in our project is to map reconstructions of film-viewing experience within a digital framework. We want this analytical framework to feature vocabulary that is as generic as possible in order to accommodate different strands of film and media studies, allowing all kinds of film and media scholars concerned with audio-visual material to ground their studies in empirical reconstructions of audio-visual composition. This means the innovative aspect of this line of work lies in a consistent as well as open process of data modelling that – in a best-case scenario – can serve as a fundamental starting point not only for our film analytical studies, but also for respective research projects in the field.

44 With regard to the variety of different forms of audio-visual material as well as our aim of conducting comparative corpus studies, we had to address the question of how to determine the parameters on which to analyze the flow of images and the granularity with which they are annotated.[12] Moreover, the annotation process – involving a group of multiple annotators – created the need for further systematic operationalization. These methodological considerations resulted in turning the focus on three basic requirements:

1. Creating a modularly structured vocabulary that is both grounded in a broad methodological film-analytical consensus and applicable with regard to specific theories on the aesthetic experience of audio-visual images (see Section 1.1).
2. Setting up a mode of description that is defined, operationalized and condensed to a degree that allows for the – to the greatest possible extent – impartial annotation of audio-visual corpora carried out by a group of trained annotators.
3. Defining the terms and procedures in a way that is explicit enough to allow researchers coming from different theoretical backgrounds to relate their approach critically to the analytic data.

45 With regard to the first point, we selected the vocabulary either directly from the broad and manifold spectrum of approaches to film analysis that focus on and describe formal elements of audio-visual composition or transformed film-analytical key concepts into annotatable keyword systematics.[13]

46 In order to meet the requirements outlined above, we chose to arrange our annotation vocabulary within a threefold structure:

- Annotation levels (namely *Segmentation*, *Language*, *ImageComposition*, *Camera*, *Montage*, *Acoustics*, *BodilyExpressivity* and *Motifs*) are the primary categories that refer to different articulatory modes of cinematic staging principles. With regard to these macro categories, we followed upon the basic levels of the eMAEX framework (see Section 1.2).[14]
- For each level we defined a multitude of annotation types in order to systematically differentiate formal

principles within these levels (e.g. *CameraMovementDirection* or *CameraMovementSpeed*). Thus, we identified analytical subdimensions that are restricted enough to provide a set of predefined terms or in exceptional cases a focussed description in free text format. Together these various types provide a basic and many-layered impression of the overarching categories. Camera is thus described as *Recording/Playback Speed, Depth Of Field, Defocus, Camera Movement Unit, Camera Movement Type, Camera Movement Speed, Camera Movement Direction, Camera Angle, Camera Angle Canted, Camera Angle, Vertical Positioning, Lens*. Needless to say, such an approach can never provide a complete description of all stylistic nuances but is rather designed to grasp central dynamics.

- For each annotation type, we defined annotation values, determining the vocabulary that can be annotated. This third step provides the basis for the actual annotation process, that consists of linking these values to a specific time increment of a given video file based on its timecode. For example, in the case of *CameraMovementDirection* the values describe the basic directions a camera can move to (e.g. left, right, up, down, towards and away but also circle, canted and undirected for more complex movement patterns). Thereby the free-text-descriptions in the original eMAEX framework are replaced by data sets drawing on a (mostly) controlled vocabulary. Annotation types are based on different internal logics with regard to how they organize values. Some feature values in reference to an ordinal scale (e.g. *Field Size*), others follow the logic of nominal scales with no hierarchical order. This difference is especially of importance with regard to data visualization (see Section 2.3). Also, different annotation types within are based on different principles with regard to how and what vocabulary can be entered, leaving us with the possibility to work with free text when necessary.[15]

47 The resulting annotation routine encompasses 8 annotation levels, 78 annotation types and 502 annotation values. All of these different descriptive dimensions are published under creative commons license (ada.ontology.org), each accompanied by a short definition explaining the use of the vocabulary. The vocabulary is regularly updated as well as our set of annotations (including the 22.000 annotations generated over the course of our pilot case study on *The Company Men*). So far the following films have been annotated: the feature films *The Company Men* (John Wells, USA 2010) and *The Big Short* (Adam McKay, USA 2015) as well as the documentaries *The Inside Job* (Charles Ferguson, USA 2010) and *Capitalism: A Love Story* (Michael Moore, USA 2009), furthermore a selection of features from the German TV News-Broadcast Tagesschau and the web clips *Occupy Wall Street* (Sem Maltsev, USA 2011) and *Where Do We Go From Here? Occupy Wall St.* (Ed David, USA 2011). All annotations are provided under creative commons licence at https://projectada.github.io and can be browsed and queried through the annotation explorer web app: http://ada.filmontology.org/explorer/.

48 Our vocabulary has been modelled as a machine-readable semantic data structure that is essential for intertwining manual and (semi-)automatic annotations, enabling the future application of machine learning, data evaluation, and potential exchange of annotations between different researchers. The respective machine-readable data ontology was set up by Henning Agt-Rickauer. It stores all values possible within our annotation framework – not as a mere unstructured text, but instead modelling the relations and dependencies within all annotation values, types and levels (e.g. the interior logic of ordinal scales) with technologies of the semantic web.

49

> 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. 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. [Agt-Rickauer et al. 2018b]

Drawing on this interconnection of semantic technologies and tools for film analysis

> “the project aims to provide an ontology for film-analytical studies complemented by a video annotation software adapted for authoring and publishing Linked Open Data by non-experts [in the field of semantic web technologies] [Agt-Rickauer et al. 2018a]

Hence, the adjusted version of the open-source software Advene can serve as a user-friendly interface, offering the advantages of semantically structured datasets to researchers without further programming skills. Annotating based on this semantic data ontology also provides the possibility to search and compare analyses based on complex queries, visualizations and other tools.

Most importantly, by providing an ontological data structure a) drawing on film-analytical key concepts as well as b) featuring short definitions with regard to the whole controlled annotation vocabulary, we designed the video-annotation framework as open as possible with regard to film-analytical studies based on different theoretical frameworks and epistemologies – creating the possibility for other researchers to relate their annotations to ours and vice versa.[16] Thus empirical data for the analysis of audio-visual material can potentially be exchanged between projects with different research questions, theoretical backgrounds or even languages.[17] 50

2.2 Challenge 2: Setting up a systematic video-annotation routine

With regard to the systematization of the video annotation, the goal was not to achieve a complete congruence of singular annotations but rather an intersubjective identification of the progression of several annotations. The dynamics of audiovisual expressivity unfold their meaning – like a melody – in the dynamic progressions over time and not in the attribution of singular values at static points in time. For example, a comparison of scene analyses by different annotators made us see that a common pattern of increasingly closer field sizes was clearly detectable even when singular decisions such as between medium closeup or shoulder closeup diverged. This fuzziness of individual annotation values is not an artefact of manual annotation but grounded in the nature of the object of study, since the film analytical concepts do not designate discrete entities – any computational distinction between, say, a close-up and a shoulder close-up is purely arbitrary. Spoken within the music metaphor: the aim of the systematization is the common recognition of a melody and not primarily that of individual notes. 51

In order to advance the annotation with our film-analytical vocabulary from a proof-of-concept state to an application-oriented methodological framework, we set up a systematic routine with the need to operationalize the process in a number of ways. In the development phase of this vocabulary, we regularly met with our expert annotators to ensure a high degree of (intersubjective) consistency of our annotation data as well as having immediate feedback on the practicality of the various concepts. Based on these repeated feedback loops, definitions for each level, type and value were acquired that guide analytical decisions during annotations. These definition texts constitute the first results of our ongoing research process. Additional insights are fixated in a technical and in a methodical manual that will be published and translated at the end of the project. 52

Furthermore, since annotating manually within the presented framework is very labour intensive – given the vast number of annotation types and values –, we had to speed up the annotation process. On the one hand, this became possible by adjusting the user interface (UI), particularly the manual aspects of video annotation (e.g. the input of values, but also the evaluation and correction of annotated values). On the other hand, the manifold description levels can be streamlined in order to analyze a larger amount of films for corpus studies. Finally – as mentioned in the introduction to this section –, we make use of some (semi-)automatic annotations based on digital tools, the possibilities of which are still to be evaluated further and implemented. 53

Regarding the UI, we optimized the manual input of controlled vocabulary through autocomplete and short keys. This allows for a quick enrichment of preexisting segments (e.g. shots or music pieces), especially with regard to annotation types with reduced input options. Combined with short key controls of the video player and switching between annotations, the annotation speed was increased. To offer the annotators an improved overview of the various annotation types (displayed as tracks within Advene's timeline-view), we developed a color code for all types, grouping them visually according to the respective annotation levels. 54

Also, over the course of the annotation of *The Company Men*, we partly re-oriented the annotation process towards longer segments.[18] In turn, we implemented the combination of multiple values for annotation types that can change even within small segments. Since the field size, for example, can change through camera movements or movements 55

within the image, there was a need to implement a simple syntax of combining values within a segment. Therefore, we have established the syntax values [TO] to indicate developments between values (e.g. “closeup,[TO],medium shot,” indicating that all field sizes between these values are passed) and [VS] to indicate two simultaneous, conflicting dynamics within a segment (for example the expressive body language for a shot encompassing two figures can be annotated as “happy,[VS],sad”).

But even with these simplifications, annotating large amounts of audio-visual material on about eighty different description levels remains an extremely time-consuming task. Since the importance of formal levels can vary within the context of different guiding questions and theories, we selected a subset of annotation types for the running AdA-project that allows film scholars to identify the basic compositional principles of a film based on a reduced set of twenty annotation types (with the possibility of annotating additional types for a selection of key scenes) from all annotation levels. 56

As mentioned above, (semi-)automatic video analysis tools offer another promising way of significantly speeding up video-annotation processes. Based on the interactions of expert annotators and automated tools (with the level of human interference depending on the amount of training the respective tool demands), an increasing proportion of manual annotation work on some (but not all) types could be (semi-)automatized in the long run. On the other hand, this bears the risk that automatically detectable traits are overemphasized for pragmatic reasons, thus creating a bias for easily created quantifiable metadata. In order to avoid respective data distortions, we decided to set up an analysis framework based on manual description first – and then evaluate the potentials for detector implementation step-by-step. 57

Following this approach, we want to briefly discuss a few automated features that have been integrated in the annotation process so far, with others still being evaluated or in a developmental stage. Already implemented in Advene was an open-source shot detection and a graphic user interface to correct the results manually (see Figure 6). 58

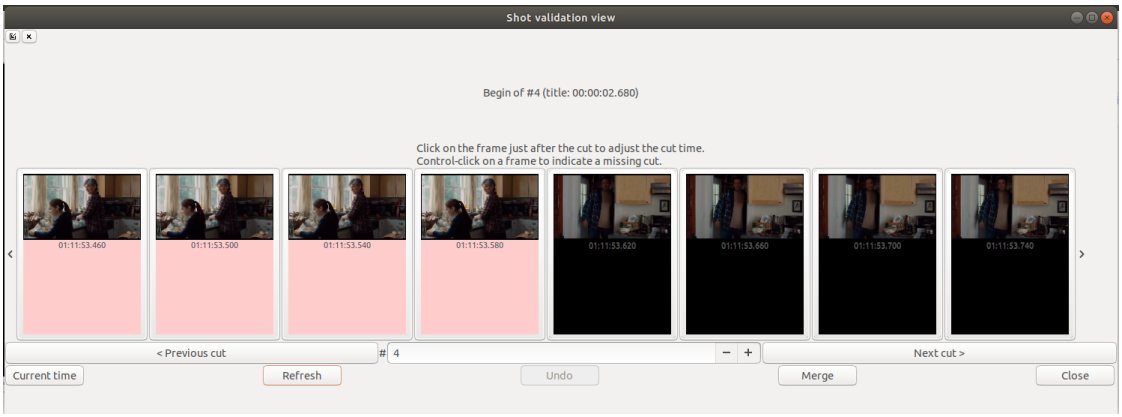

Figure 6. Advene’s shot-validation view.

In addition, we use a second proprietary detector to analyze shot transitions which achieved much better results on fades, wipes and other continuous transitions. With approximately 1.500 frame-precise annotated shots over the course of a film with a two-hour runtime, the feature of automatically detecting cuts and shot transitions can be seen as a huge advance in comparison to manual work – especially given the essential function of the “shot” segment as basic micro segmentation with regard to many other annotation types. 59

Another already widely used automated tool depicts the general volume of a video file’s audio track as a soundwave, allowing to quickly grasp which sections of a film are especially calm or loud and where sudden changes in volume occur (see Figure 7) – a feature we use, for example, in order to quickly identify peaks in volume, such as the prominent use of loud music or sudden noises like a gunshot. 60

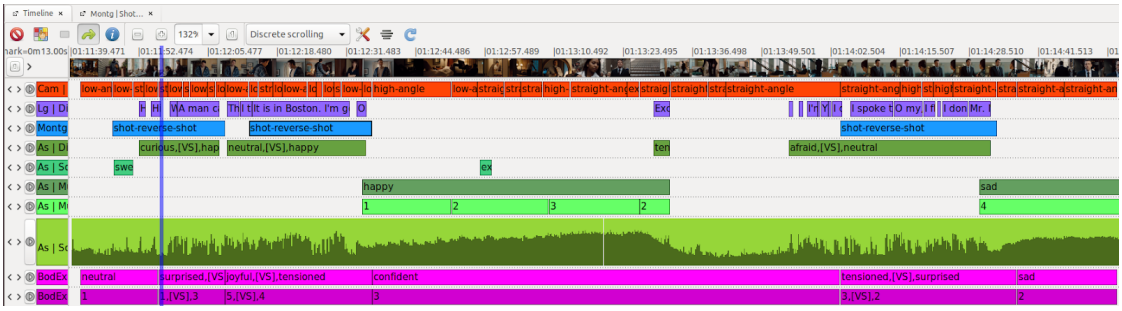

Figure 7. Waveform implementation in Advene's timeline-view.

Another field with a high potential of automatization that is still in the state of evaluation concerns written and spoken words. Especially regarding audio-visual material where subtitles[19] are not available, automatic language recognition can provide a solid base for manual corrections or – at least – a general indication where language appears and where not. Since our project features non-English documentaries and German television news, the discrepancy between English and other speech-to-text tools became evident. Regardless of the technological base, a human correction of transcoded dialogue is currently still indispensable. In light of our goal to produce an open and free framework, we refrained from using language processing from Amazon or Google which might provide better results. 61

Other areas for applying computational analysis that are currently under development are color detection,[20] automatic detection of aspect ratio and optical flow-analysis. In addition, there have been convincing attempts at using algorithms for face detection in order to automatically detect field sizes [Arnold and Tilton 2019]. But however promising the values for precision and recall may be, they are still in a range that is more feasible to statistical abstractions and distant viewing than to the needs of a precise qualitative reconstruction. Therefore, they would still need a lot of manual corrections in order to obtain continuous correct field-size-annotations for a full-length film. So far, the automatically generated annotations did in many other cases not comply with the manual annotations based on human perception in an acceptable margin. Future research will have to show which of these tools can be adjusted to the requirements of a qualitative "close viewing" to a degree that makes manual correction obsolete or manageable. 62

2.3 Challenge 3: Developing visualizations for patterns of audio-visual composition

After discussing the strategies for improving the process of entering and correcting annotations, we want to discuss the challenge that arises from annotating a film extensively: How to work with this complex set of data, without leaving the "reading" of data solely to algorithms and statistics? For an encompassing analytical approach, the question of assembling and relating annotations becomes relevant. As mentioned before, the film The Company Men was annotated[21] across 66 different annotation types which led to a data set of approximately 22.000[22] annotations for a single feature film. 63

This amount of metadata of course raises the question of "readability" which we will be discussing in the following by describing our visualization efforts. The visualization of our complex data sets can produce immediate insight into a composition and provide the involved film scholars with the possibility of guiding software-based searches for recurring compositional patterns. 64

Referring to the arrangement of a timeline with x and y axis, such a "reading" of annotations has a horizontal and a vertical dimension. In this context, "horizontal" refers to the temporal succession of annotations, whereas "vertical" refers to the synchronicity of annotations (i.e. values) assigned to different annotation types. In terms of this basic distinction, different forms of visualization with different advantages are possible and necessary for specific purposes. Thus, not a single visualization paradigm (e.g. multilayered timeline, histogram of a single type, etc.) can be singled out. In turn, a toolbox of different ways to enter and read annotation data has to be provided, that varies according to the respective research interest and theoretical framework. For example, data may be presented in a table or a timeline.[23] 65

The visualization of values in a timeline allows for a quick and intuitive understanding of the length of single annotations as well as the rhythm and the compositional patterns they form together with other annotations. With regard to our research on audio-visual rhetorics of affect and the comparative analysis of compositional patterns, this contextualization of single values is crucial.[24] 66

In our joint efforts with Olivier Aubert, a visualization feature was developed that is precisely adapted to our research project's comparative scope: the AdA-timeline (demo: https://olivieraubert.net/hpi/timeline.html).[25] The respective diagrams (see Figure 8) can be directly generated with Advene, offering the possibility to instantly adapt to changing or developing sets of annotations. 67

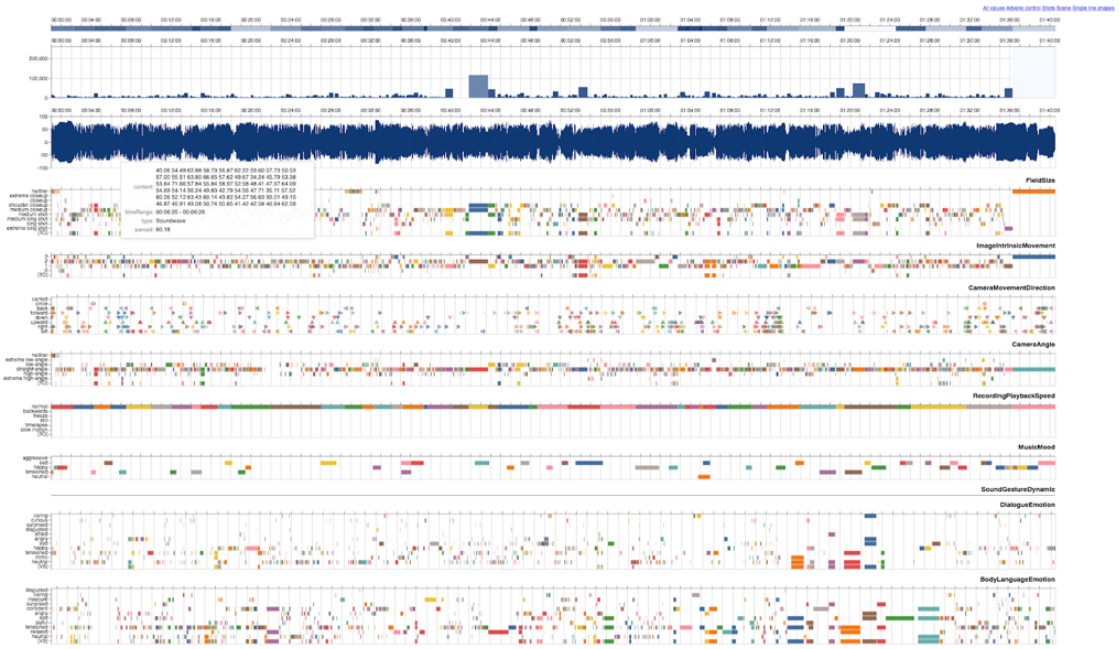

Figure 8. Screenshot of the AdA-timeline generated with Advene.

The basic idea behind this mode of visualization can be described as making a selection of annotation types readable like a musical score of an orchestra piece, displaying audio-visual rhythms as graphical patterns. Here we draw on film scores from theorists like Sergej Eisenstein [Eisenstein 2006], but also newer examples developed in projects like "Digital Formalism" or "Cinemetrics" as well as on scores Jan-Hendrik Bakels developed in his book on audio-visual rhythms [Bakels 2017]. 68

For example, the dynamics of *FieldSize* in Figure 9 can be grasped visually; extrema are more easily and quickly detected than in a single-line depiction. 69

Figure 9. Single line of Advene timeline view vs. multiple lines in the AdA-timeline.

The AdA-timeline (as seen in Figure 8) features at the top a timeline of the whole film (see Figure 10), indicating different scenes in various color shades. By marking a segment in this timeline, it is possible to zoom into the respective subsegment – switching from a micro-perspective of a few seconds to an overview of the whole film within an instance. Below this zoom bar is a histogram of shot length (similar to Cinemetrics' visualizations) displayed for the currently 70

selected segment.

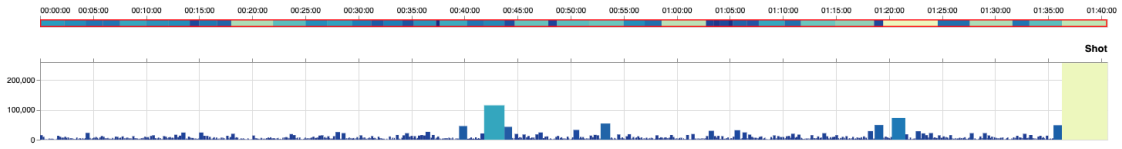

Figure 10. Screenshot of histogram in the AdA-timeline.

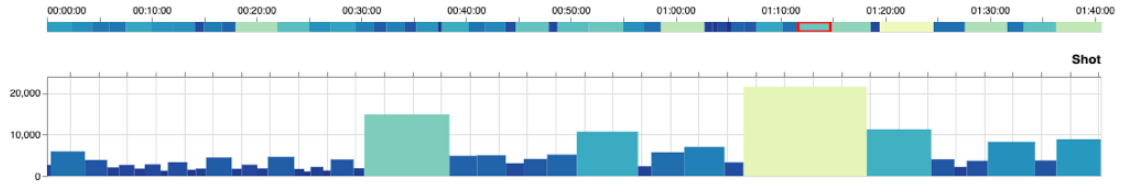

Figure 11. Screenshot of zoom bar in the AdA-timeline.

This offers the possibility to quickly navigate across different points within the running time of the film at hand as well as across different annotation levels, types and the respective values. Furthermore, it is possible to choose and display only a selection of annotation types within the diagram, to change their color palettes, to switch between different representations (such as single line, horizontal bar graph and, in some types, histograms and waveforms), and to connect the browser-based visualization with an embedded video player, so that by clicking on an annotation, the respective segment can be watched. 71

Another way to navigate and/or filter the diagram could be to combine it with query interfaces, identifying scenes across the corpus based on complex sets of search parameters (e.g. a search for all segments where closeups occur while sad music is playing). Currently, Henning Agt-Rickauer is developing such an interface in cooperation with Joscha Jäger – the annotation explorer,[26] which is based on FrameTrail – for the AdA-project.[27] In combination with graphical evaluations of visualizations and query tools, an image search developed by Christian Hentschel is another useful tool for a comparative navigation of the corpus in regard to motif studies and the analysis of the modulation of affective profiles. 72

In this section's last subchapter, we will take a second and last look into the scene from *The Company Men* described within the eMAEX framework above (see Section 1.3) in order to present a short use case for the methodology outlined in this section. 73

2.4 Exemplary Analysis II: Studying The Company Men based on visualized video-annotation data

Over the course of this second look at our exemplary scene, we want to show how the outlined affective parcours (from excitement, joyful expectation, and bafflement to sad isolation), as well as the described EMUs and their figurations, can be detected and substantiated within a bottom-up perspective by retracting compositional patterns from our annotation data or its visualization. The initial and decisive step, i.e. the segmentation of the scene into four EMUs, can already be retraced within a brief overview of a selection of annotation types and the respective annotations. Figure 11 shows an excerpt from the Advene timeline view, depicting annotations according to annotation types and in their temporal expansion. 74

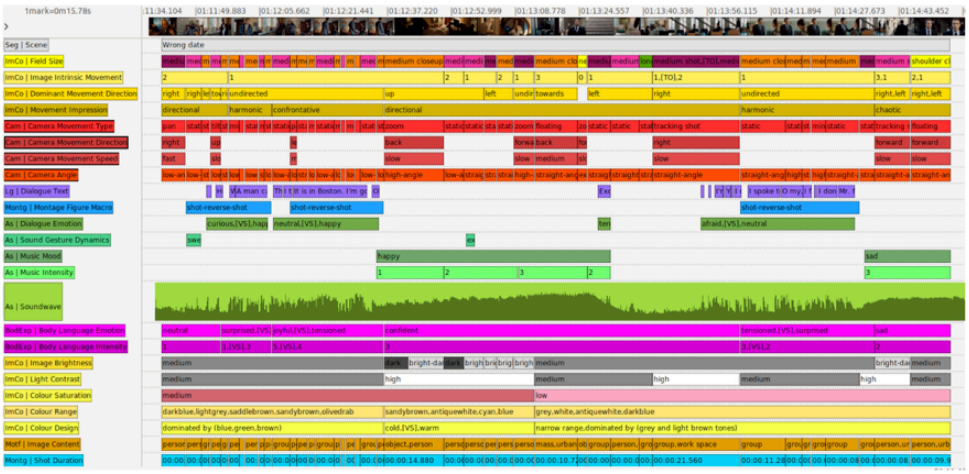

Figure 12. Advene timeline view of the scene “Wrong Date” TC 01:11:41–01:14:56. *The Company Men*. John Wells, USA 2010.

Already at first glance, the on- and offset of music (green annotation types *MusicMood* and *MusicIntensity* in the middle of the Advene timeline view), as well as the dialogue on- and offset and the use of shot-reverse shot-montage (purple annotation type *DialogueText* and blue annotation type *MontageFigureMacro*) indicate a clear structuring of the scene into four parts based on rhythmic patterns (see Figure 12).[28] 75

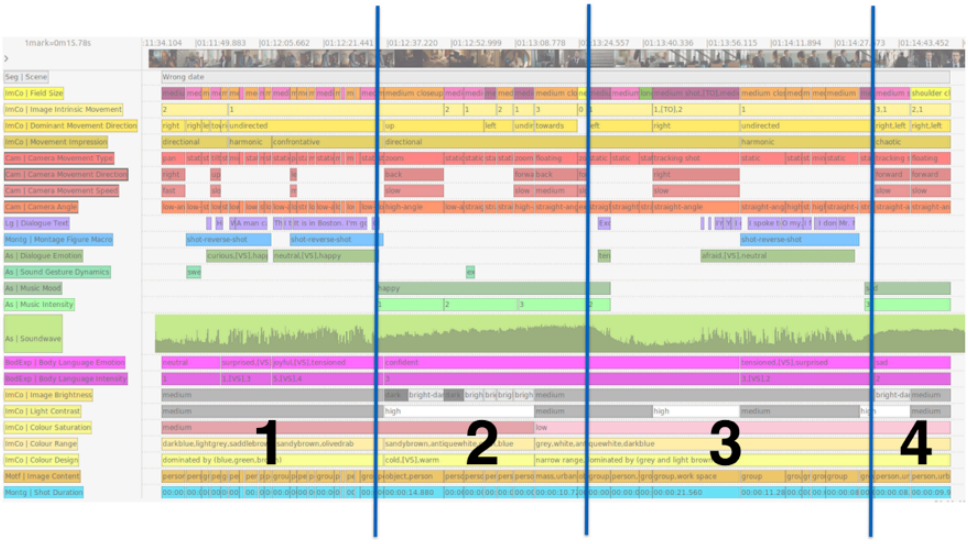

Figure 13. Segmentation of the scene “Wrong Date” on the basis of annotations TC 01:11:41–01:14:56. *The Company Men*. John Wells, USA 2010.

In the following, we will take a closer look at the four EMUs and explain how we can build our film-analytical claims bottom-up, drawing on our annotations. 76

Figure 14. The first expressive movement of the scene "Wrong Date" TC 01:11:41–01:14:56. *The Company Men*. John Wells, USA 2010.

We described the first EMU – which we qualified as an abruptly emerging, then continually rising energy – as a figuration of suddenly increasing vitality that is predominantly shaped through bodily expressivity,[29] montage[30] and the acoustic composition[31] of the segment. The drastic change in the acting style can be observed in the annotation type *BodyLanguageIntensity*[32] (see Figure 13) in which the intensity is rated on a scale from 1 (low) to 5 (high intensity). Whereas the first shot is rated low (1) and in the second shot a low (1) and a medium (3) intensity are confronted, the intensity rises again – rather abruptly – in the third shot with the contrast of 5 and 4, i.e. a discussion between the protagonist and his wife. This gradual rise of *BodyLanguageIntensity* overall, as well as the sudden significant increase – which can be ascribed to Bobby's (Ben Affleck) *BodyLanguageIntensity* – is accompanied by an increase in volume that can be traced in the waveform depiction. In this second half, the *ShotDuration*[33] shows a particular rhythm: several short shots with closer field sizes showing Bobby or his wife are followed by a longer medium shot of the bedroom (each with Bobby's wife facing the camera in the background). This specific pattern repeats and gradually accelerates during this segment – with a slight deceleration at the end of the scene.

77

Figure 15. The second expressive movement of the scene "Wrong Date" TC 01:11:41–01:14:56. *The Company Men*. John Wells, USA 2010.

We described the second EMU – which we qualified as vivid directedness – as a swelling figuration. Based on the

78

annotations it becomes apparent how the music is dominant in this regard: i.e. its increasing volume which can be deduced again from the waveform and that correlates with an increasing *MusicIntensity*.[34] The latter was annotated as rising from 1 to 3 (on a scale from 1 to 5). Furthermore, this music piece was qualified by our annotators as upbeat, happy music.

79 In addition, the swelling dynamics of the music piece interact with other annotation types. As the bottom of the Advene timeline view in Figure 14 shows, the image brightens[35] continually from dark shots (black) over bright-dark shots (contrast of bright and dark parts of the image depicted in light grey annotations) to consistently medium light sequences (in medium grey annotations), while at the same time the *ColourRange*[36] used by the film changes distinctly. Furthermore, the swelling movement is constituted through an increasingly dynamic cinematography, identifiable by the use of faster camera movements at the end (see the red annotation type *CameraMovementSpeed*[37] that changes from slow to medium).

80 In the last part of the segment, the camera interacts with the image intrinsic movement (see the yellow annotation type *ImageIntrinsicMovement*[38] that describes the perceived degree of movement within a shot on a scale from 0 to 3; in this segment, the dynamicity varies between 1 and 2 before reaching its climax with 3 shortly before the end of the sequence). Overall, the image intrinsic movement thus shows a course of increasing intensity, that comes to an abrupt halt when the last shot of the segment depicts in an extreme low angle a static shot of the company building that was qualified as a 0.

81 We can summarize this second expressive movement as a swelling movement, where the music intensity, the camera movement speed, as well as the image intrinsic movement jointly rise over the course of the segment. The brightening of the image, the upbeat music, as well as the facial expression together with the directedness of the camera movement and the general vividness of the image shape this swelling as a vivid directedness, finding its finale in front of the seemingly gigantic company building (extreme low angle).

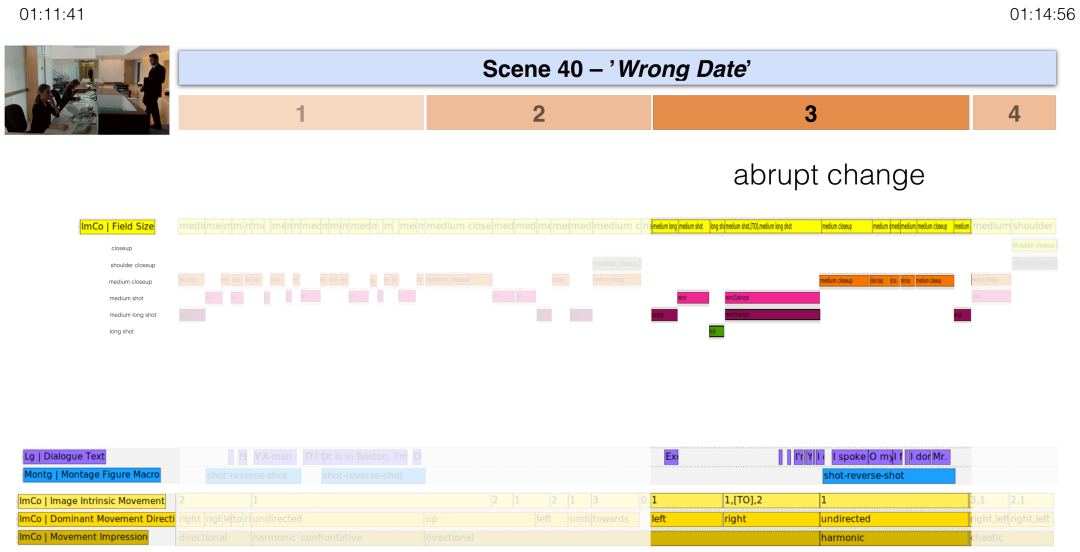

Figure 16. The third expressive movement of the scene "Wrong Date" TC 01:11:41–01:14:56. *The Company Men*. John Wells, USA 2010.

82 The third expressive movement – which we qualified as vividness put abruptly to rest – stages a figuration of abrupt change. Within the *FieldSize*[39] annotations, a shifting dynamic between closeness and abrupt distancing can be retraced. This happens twice in this EMU: a first instance briefly in the beginning, and a second more emphasized one in the last shot. At the first instance (see dark purple annotation in Figure 15), a medium long shot of Bobby standing at

the reception desk is followed by a closer medium shot of him sitting down (pink annotation). By again jumping away in the third shot (green annotation), the cinematography creates a tension.

This pattern is repeated in a rather emphasized manner and prolonged at the second instance towards the end of the EMU: The camera approaches the conversation between the protagonist and the assistant in a series of medium closeups, before returning abruptly to a medium long shot in the very last shot of the segment (see Figure 15). This second distancing movement is accompanied by a sudden silence on the level of dialogue (see the absence of purple annotations in the type *DialogueText*[40] in Figure 15). We qualify this moment of irritation, the abrupt change by means of a distancing movement accompanied by a sudden silence, as the experience of vividness put abruptly to rest. 83

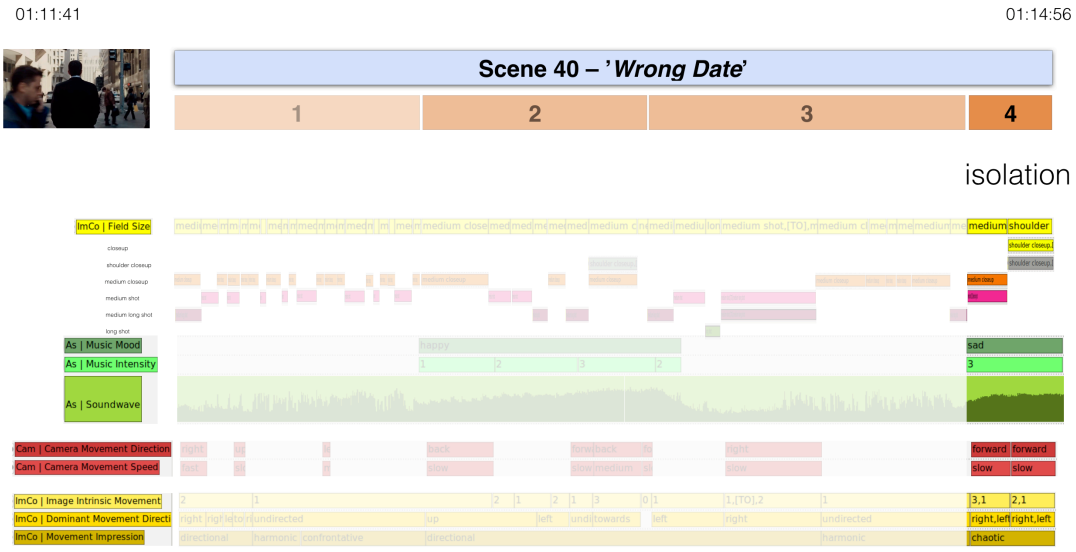

Figure 17. The fourth expressive movement of the scene "Wrong Date" TC 01:11:41–01:14:56. *The Company Men*. John Wells, USA 2010.

The fourth and last EMU of this scene – which we qualified as an experience of slowly being isolated – is identified as a figuration of an intensifying deceleration and contraction which can be substantiated as follows: Reversing the distancing movement of the last segment, this expressive movement is characterized by a gesture of approaching, realized through montage and camera movement. The camera approaches the protagonist until we see him in a shoulder closeup (the closest field size used in this scene; see Figure 16). On the acoustic level, music sets in – which is qualified as sad and intense. The observed movement dynamics within the frame at this moment are especially interesting: Two separate movement intensities are identified, so that two separate values are given and related via a [VS] syntax element. This derives from the harsh contrast between the static protagonist in the image centre and his dynamic surrounding (various moving passersby). Whereas the interplay of moving image parts was perceived before as directed or harmonic, it is now classified as chaotic, thus changing the underlying impression. Overall, we can describe this movement figuration of a camera closing in on the protagonist's tensionless face accompanied by sad and intense music and with a diverging movement dynamicity within the frame as figuration of an intensifying deceleration and contraction. 84

In summary, the scene as a whole creates an image of disappointment, realized by the succession of an abruptly emerging, then continually rising energy and vivid directedness, over a constricted vividness that is abruptly put to rest, before slowly retreating into isolation. This affective course can, as we hope to have shown, easily be identified on the basis of the compositional patterns that appear in the structured annotation of the scene. This bottom-up perspective enables even scholars not trained in qualitative description of expressive movements to identify the respective patterns. 85

The systematic description of these audiovisual patterns is no end in itself, but should enable us to search for similar 86

patterns in the wider corpus of films. Drawing on the presented analysis, we can hypothesize that scenes of experiencing disappointment are recurring throughout our corpus. We can now condense the observed pattern in our annotation data to some core traits: a shift from “happy” to “sad” music mood[41] (with a pause in-between); closer field sizes at the end and a high image intrinsic movement correlating with the annotation of happy music. In Figure 17, a responsive and zoomable timeline view (see Section 2.3) of the scene (titled) “Wrong Date” shows some of these key traits. For example, a blue-coloured bar in the annotation type *MusicMood* at the bottom visualizing the value “happy” is followed by a short segment without annotation, then a red-coloured bar visualizing the value “sad.”

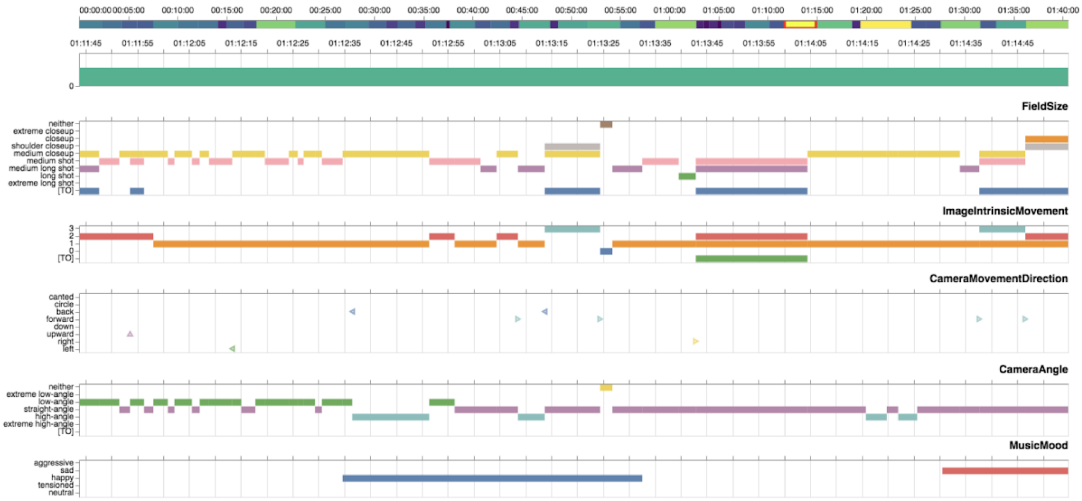

Figure 18. Reduced overview of the scene “Wrong Date” TC 01:11:41–01:14:56. *The Company Men*. John Wells, USA 2010. AdA-timeline.

Our growing annotation database allows us to query for scenes that meet these criteria. In Figure 18 below, you can see a graphical comparison with a second scene from *The Company Men*. The scene “Bobby’s last paycheck” (bottom) occurs earlier in the film, meeting the same criteria regarding the happy-sad change in music as well as the approaching camera at the end of the scene. Synchronously with the annotated *happy* music you see as in the other scene a high intensity (value: “3”) of *ImageIntrinsicMovement* as well as more (*shoulder*) *closeups* in the segment with *sad MusicMood*. In our exemplary scene we saw that this pattern was an indicator and element of the scene’s affective parcours (of excitement, joyful expectation, bafflement to sad isolation). A further in-depth analysis of this second scene would show that it does indeed stage – a slightly varied – image of disappointment.

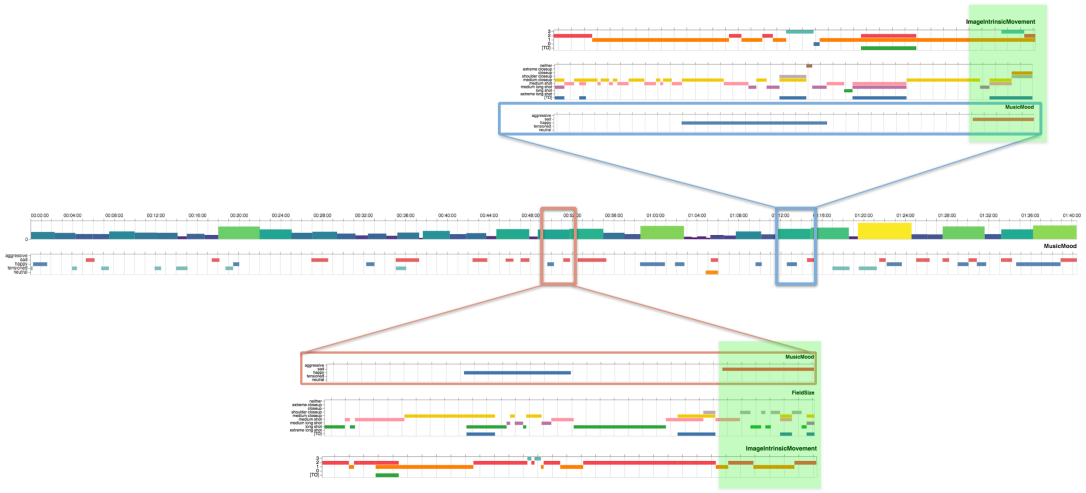

Figure 19. Comparison of two scenes in *The Company Men*. John Wells, USA 2010.

This analysis so far has shown that we can conduct a detailed analysis of spatio-temporal dynamics in moving images on the micro level based upon empirical data from the annotation of a largely controlled vocabulary. Over the course of this exemplary study, we hope to have shown how to determine key characteristics of single analyses and search for these characteristics within the film (and throughout a larger corpus) in order to find similar scenes. This enables us to identify recurring affective profiles within a corpus and thus allows us to make detailed and empirically-based claims about larger groups of films – focusing on the dynamic unfolding of scenes and their experiential quality. 88

3. Matching Computational Analysis and Human Experience – Conclusion and Perspectives

In this article we have focused on the requirements, challenges and proposed solutions we have identified within the research goal of matching computational means of analysis and interpretation with the theoretically deduced primacy of experience in media reception. The sketched-out film analytical framework is no one-fits-all-solution for analyzing audiovisual media, and there are still many questions unanswered – and even not yet asked. But we hope that this is also read as a contribution to the ongoing development of Digital Humanities methodologies which are – at least with regard to audio-visual images – still largely in a phase of fast development. 89

Apart from the concrete hypotheses, measures and findings that we have presented, we want to outline in this conclusion the underlying epistemological and disciplinary principles that are of importance to us and that also resonate in some recent debates about the directions that the field of Digital Humanities is taking: 90

The dichotomy between the close and the distant, the qualitative and the quantitative – as well as between the digital and the analog, incidentally – should not be considered as mutually excluding and framed in a discourse that emphasizes deficiency. We rather have to find means to foreground the inherent hybridity and scalability that works within the Digital Humanities and make it productive in as many ways as we can [Fickers 2018]. It is thus important to us to emphasize the circumstance that this kind of research program is not proposed as a substitution to established methods and frameworks, but rather as a promising addition to the methodological toolboxes of the different disciplines. 91

It has also been established quite often, that the hybridity of the Digital Humanities necessitates (at least) two translation efforts and literacies [Jones and Hafner 2012]: One concerns humanities scholar who may not have to become a programmer but at least has to develop a basic understanding of how codes, databases and interfaces work. But the other translation is equally important: The possibilities of computation are highly problematic from a techno-political 92

point of view, as long as they are not viewed from specific theoretical perspectives and research questions that are derived from concrete problems of knowledge and understanding. In order to relate aesthetic theory of embodied perception and methods from computational sciences, a common ground has to be found that deviates from the standards of the involved disciplines and always runs the risk to seem incompatible with conventional self-understandings – a predicament that a lot of inter- and transdisciplinary research has to face.

93 In our case, we proceeded from a theoretical concept of embodied experience and expressive movement patterns in audio-visual images. We were confronted with the task to translate this into a data structure that was accessible to computational processing. One important side effect of this and other frameworks is that it encourages collaborative research – annotating together, re-using analytical data with new questions – even though a lack of universal file formats and standard software still makes the exchange and implementation in the scholarly communities difficult. We followed the idea that ways of grasping compositional patterns in fine-grained analytical reflection certainly exceed human capacities when they are applied to larger corpora, but that this could be achieved by combining a systematic, theory-guided production of large amounts of data. This implied a different look at data as compared to previous, digital and non-digital modes of data collection in film and media studies, which did not start from the maximum of fine-grained analytical access but from a minimum of easily quantifiable features [Salt 1983]. It also implied a changing perspective on computational methods since it emphasizes a holistic consideration of formal structures within audio-visual material that overarches the enterprise of formalizing any individual feature for algorithmic detection. (So far, our experience in the implementation of automatic feature detection has highlighted the need for further development in this field, but we are convinced that this is a question of time and appropriate training data that ideally also takes into account manual annotations that are collected with a view towards aesthetic structures. The flipside of this assessment is the fact that the time effort of largely manual annotations and the training of expert annotators is still a disadvantage considering the average infrastructure of research projects.)

94 We want to propose our Data Model as a possible framework for such a holistic consideration that not only provides a modular, structured coding system for the many levels of audio-visual analysis, but also offers a view on connecting these levels in order to achieve the further application of methods like search algorithms and machine learning. One future direction of this research lies in the further use of manually gathered annotation data to train algorithms not only in the detection of single features but in the identification of recurring patterns. The other – associated – field of application is the further development of standardized visualizations and the implementation of complex cross-modal queries.

95 The basic framework provided by the AdA filmontology aims at providing a first encompassing data structure for the various stylistic levels of cinematic expressivity – with all the advantages and disadvantages that come with such an ambitious ‘global’ approach. We still hope that the general steps that we tried to sketch out over the course of this article can serve as orientation for similar endeavors in other disciplines of performative, time-based arts and media. These programmatic steps involve identifying possible starting points with regard to addressing intersubjective bases of experience, setting up a systemized machine-readable vocabulary addressing these bases, making use of visualizations and computational methods in order to identify complex, recurring patterns. With regard to these epistemological steps, temporality and patterning could serve as a common denominator for the larger field of integrating digital tools and the study of performative, time-based arts and media.

96 With these preliminary results, we hope to give new impulses for the nexus between film analytical research and the implementation of digital tools based on a machine- and human-readable defined vocabulary. We started from the finding that algorithms are not profound in reconstructing the bases of experiential qualities like feelings. But when confronted with a research question that takes a look at a larger corpus of audio-visual material, one soon finds that it is not strictly speaking possible to objectify and empirically compare these experiential qualities apart from the compositional patterns in which they are grounded. It is this diversion via setting up a systemized machine-readable vocabulary addressing the intersubjective bases of film-viewing on the level of modeled data, that makes it possible to match algorithmically interpreted and generated annotations with research interested in the experiential qualities of film.

Notes

[1] Burdick et al. write: “Rather than pitting distant reading against close reading, what we are seeing is the emergence of new conjunctions between the macro and the micro, general surface trends and deep hermeneutic inquiry, the global view from above and the local view on the ground” [Burdick et al. 2016, 39]. In line with this reasoning, our interest is to investigate “new conjunctions between the macro and the micro” with a focus on questions of aesthetic experience.

[2] Of course every analytical approach to arts and media includes forms of abstraction due to the mere fact of conceptualization and verbalization, but we would like to give weight to the difference between abstractions that aim at processes of embodied and situated reception as their primary data and abstractions that treat the coded data-sets as their primary object [Drucker 2016].

[3] Members of the AdA-project are Jan-Hendrik Bakels, Thomas Scherer, Jasper Stratil (Freie Universität Berlin), Henning Agt-Rickauer and Christian Hentschel (Hasso-Plattner Institut) with project mentoring by Hermann Kappelhoff (Freie Universität Berlin) and Harald Sack (FIZ Karlsruhe/ Hasso-Plattner Institut). Associated Members are Matthias Grotkopp (Freie Universität Berlin) and Olivier Aubert (Université de Nantes). The project is funded by the German Ministry of Education and Research (BMBF), Dez. 2016 – Nov. 2020. See also: http://www.ada.cinepoetics.fu-berlin.de/en/index.html.

[4] We draw upon a wider context of research, see e.g. Apostolidis and Mezaris 2014, Petersohn 2008, Petersohn 2009, Mashtalir and Mikhnova 2011, Szeliski 2011, Krizhevsky 2012, Russakovsky et al. 2015, Yosinski et al. 2014, Datta et al. 2006, Su et al. 2005, Zhang and Tomasi 1999, Hentschel et al. 2013, Agt and Kutsche 2013.

[5] The eMAEX (short for *electronically based media analysis of expressive movement images*) system was developed by a group of film scholars led by Hermann Kappelhoff at Freie Universität Berlin. For more information see: https://www.empirische-medienaesthetik.fu-berlin.de/en/emaex-system/emaex_kurzversion/index.html.

[6] The following exemplary study focuses on a scene from a contemporary Hollywood feature film. However, the eMAEX system has been among others applied to feature films from different genres, regions and historical periods, documentary and propaganda films, film and TV-news, short and long web video formats, animation films and advertisements: there have been studies on Hollywood war films from the 1940s until today [Kappelhoff 2018a] [Kappelhoff et al. 2013] [Scherer et al. 2014], contemporary German arthouse cinema [Schmitt 2020], Hollywood auteur cinema (contemporary [Bakels 2017]; paranoia cinema from the 70s [Lehmann 2017]), 1940s screwball comedies [Greifenstein 2020], French silent films [Berger 2019], film noir [Müller and Kappelhoff 2018], documentaries (climate change [Grotkopp 2017]; Iraq war documentaries [Pogodda 2018]); German tv news [Müller and Kappelhoff 2018], American newsreels from the 1940s [Gaertner 2016], commercial advertisement [Schmitt 2020], political advertisement from Germany and Poland [Horst 2018], as well as animation film ([Kappelhoff 2018b], [Hochschild (in preparation)]) and web video formats (YouTube vlog [Stratil 2020]; online activism [Bakels et. al in preparation]).

[7] In order to segment a film into scenes, at least three expert annotators with prior film analytical expertise segment each film into scenes. These protocols are then merged. Guiding for this segmentation are narrative clusters (indicated through leaps in diegetic time or a change of settings), thematic and discourse units as well as audio-visual units (e.g. with fade-to-blacks or music usage as prominent markers, but also more complex changes in the staging mode, e.g. from a fast-paced action-sequence to a shot-reverse-shot conversation). In other contexts, scenes are defined along paratexts from the production (screenplays) or publication (DVD/Bluray-chapters). Since these divisions are often not congruent with each other and exclude the viewer experience completely, they are not applicable to our research focus.

[8] An ideology-critical reading of the film *The Company Men* (John Wells, USA 2010) as a whole and its conservative take on gender roles and economics can be found in [Kinkle and Toscano 2011].

[9] In regard to the practical limits of the film analytical method, the turn to the field of digital video-analysis tools is closely connected with the hope for saving time: the qualitative description of audio-visual sequences (e.g. within the eMAEX framework) is extremely time consuming. Analyzing a film or video in most cases takes a multiple of their running times. And the manual annotation on a high number of formal levels within a large corpus of audio-visual material threatens to be no less time-consuming. However the emerging field of video analysis and retrieval within the computational sciences has developed a lot of tools for (semi-)automatically analyzing formal aspects of audio-visual material.

[10] See Endnote 6.

[11] Various video-annotation software in the past decades have developed different approaches to creating working environments with different emphases: From collaborative live-tagging of videos (LookAt, OpenVideoAnnotation, TRAVIS) to the depiction of automatic concept detection (VATIC), the coding of audiovisual data in the tradition of empirical social research (MAXQDA, AQUAD), real-time annotators for shot frequency and field size (Cinemetrics), the measurement of specific aspects of a film (e.g. shot length in Cinemetrics) to layer based tools for complex annotations with free text or controlled vocabularies (ADVENE, ANVIL, ELAN, VIAN). Researchers interested in complex, dynamic and

multimodal film analysis have mainly used the latter. Especially the timeline of a film, a video player and visually separated annotation layers for observations in different description levels provide the backbone of most annotation programs used in the field. These interfaces are similar to those of popular video editing programs like Adobe Premiere, Final Cut or Avid (that are also sometimes transformed into annotation programs by scholars, c.f. Jacobs and Fyfe 2016). The timeline indicates in this sense the relevance of the aspect of temporality for the analysis of audiovisual images that can be also seen governing the central principle of (temporal) segmentation put at work in our praxis of video annotation. Besides the graphically adjustable timeline especially the in-application generation of visualizations and multimedia publications were decisive factors for our initial choice for Advene.

[12] The question of segmentation is not only relevant to macro units – scenes or compositional segments that can only be grasped cross-modaly like the EMU –, but is raised by every single annotation on any level of description.

[13] As an example for film-analytical key concepts, the basic distinctions of mise-en-scene, cinematography, editing, sound that can be found in propaedeutic literature on film analysis [Bordwell and Thompson 2013] [Corrigan and White 2012] can serve as points of orientation.

[14] We slightly adapted the basic levels of the eMAEX framework: e.g. we transformed “gestures and facial expression” to *BodilyExpressivity* or added the level *Language* for the semantic dimension of written and spoken word, due to its heightened role in TV-news and some documentaries. The level *Segmentation* addresses macro-units (e.g. scenes and EMUs). The basic division into subunits (scenes and shots) is undertaken prior to splitting up the film into various packages along the scene-units to allow a synchronous annotation by multiple annotators; the segmentation into EMUs is carried out later in this revised process, based on the detailed annotation data.

[15] In some annotation types, we refer to concepts that derive from the specific films and cannot be anticipated beforehand; in other cases longer sentences are needed to describe more complex figurations that oppose the restrictions of a defined vocabulary. Also, free text can always be added for every annotation type that features fixed values in order to maintain the possibility of adding further analytical observations.

[16] The AdA filmontology aims at meeting the demands of a methodology based on the premises of expressive movement analysis but is not predetermined by our specific research question in such a way that it excludes researchers working with different assumptions. It is our hope that it can be deployed as a starting point for a toolbox that offers a basic framework for the empirical study of audio-visual composition or as a reference data set for film analytical tools.

[17] Annotation levels, annotation types and annotation values are machine readable concepts with attached labels in natural language. Thus it is possible to refer in different languages to the same concepts. So far we have implemented identifiers and definitions for each concept in English and German.

[18] To give an example for the extension of segments: we defined guidelines for identifying shot overarching color spaces instead of annotating color types for each single shot.

[19] Let alone the factor that subtitles often do not offer word-by-word accounts of the actually spoken dialogue.

[20] Yet, especially the mapping of easily understandable color descriptions like “dark blue” with detector results that provide numeric color ranges in different color systems remains a main obstacle on the path towards (semi-)automatization.

[21] We thank our student assistants and expert annotators Anton Buzal, Yvonne Pfeilschifter, João Prado, Maximilian Steck und Rebecca Zorko (all Freie Universität Berlin) for their patience, critical minds and curiosity.

[22] A scene segmentation that was first performed by three expert annotators independently and later combined, divided the whole film into 49 scenes. The shot recognition was done by two different shot detection logarithms. These results were corrected and merged by two annotators. This process led to a division of the film into 1206 individual shots. The whole film was then segmented in the individual scenes and shared within a group of four trained annotators.

[23] A table view of one single annotation type allows for an easy reading of annotation values within this type, but not for examining possible interrelations with values referring to another annotation type. Also the temporal dynamics these values reflect with regard to a larger segment will be more difficult to grasp.

[24] Regarding the contextualization of single values for example, a two second closeup within a series of close shots is to be considered significantly different with regard to viewer addressation than a twenty second closeup after a series of long shots.

[25] Please note that some browsers have problems loading this page. We recommend Firefox.

[26] http://ada.filmontology.org/explorer/

[27] See https://frametrail.org/.

[28] It is important to note that the scene's segmentation can be based solely on rhythmic patterns that indicate how the scene is staged, and not necessarily on represented content like the setting or narratively constructed diegetic time frames – even though these aspects often concur.

[29] "Bodily expressivity" is defined in the AdA filmontology as: "Expressivity of bodies that are perceived as communicating bodies (e.g. humans, animals, anthropomorphic machines). The expressivity is not understood as a speculation about an assumed subjectivity, but as perceived surface phenomena of gestures, facial expressions and postures."

[30] "Montage" is defined in the AdA filmontology as: "Staging strategies that only result from the interrelation of two or more shots. Montage refers here to the cutting of subsequent or co-occurring shots, as well as to the assemblage of sequences as temporal gestalts. The emphasis is on visual editing, sound editing is primarily annotated under 'acoustics.'"

[31] "Acoustics" is defined in the AdA filmontology as: "This level encompasses all annotation types that refer to the staging of expressive acoustic phenomena like music, sound design, or the expressive qualities of spoken language."

[32] "BodyLanguageIntensity" is defined in the AdA filmontology as: "Perceived degree of dynamicity and tension in an affective expression regarding the body language (gestures, posture, as well as facial expression) of central figures within the image. It can also involve an inward-oriented form of tension, such as repressed anger. This annotation type provides a scale for the intensity of body language. Conflicting intensities (e.g. different figures in the image or a difference between gestures and facial expressions) can be related as conflicting in the sense of a 'versus' with [VS]."

[33] "ShotDuration" is defined in the AdA filmontology as: "The temporal duration of a shot. A shot of a film is a perceivable continuous image and is bound by a 'discontinuation of the entire composition.' (Fuxjäger: Wenn Filmwissenschaftler versuchen, sich Maschinen verständlich zu machen, 2009, own translation). In this annotation type, the shot duration is stated in seconds."

[34] *MusicIntensity* is defined in the Ada filmontology as: "Perceived degree of the intensity of an (affective) expression of music, e.g. regarding volume, dynamics, instrumentation. This annotation type provides a scale for the intensity in a coherent segment of music (either a piece or a part of it)."

[35] Image Brightness is defined in the AdA filmontology as: "Perceived light intensity of a shot. This annotation type provides a rating scale for image brightness that refers to film-intrinsic variations and not to absolute values."

[36] "ColourRange" is defined in the AdA filmontology as: "The perceived range of (main) colours in a sequence. In this annotation type, for the purpose of comparability, colours have to be picked from a reduced set of colours. A description of the colour impression is combined with a hex color code of the corresponding colour value as a reference."

[37] "CameraMovementSpeed" is defined in the AdA filmontology as: "Perceived degree of the (relative) movement speed of the camera. This annotation type provides a scale for the perceived camera speed from slow to fast."

[38] "ImageIntrinsicMovement" is defined in the AdA filmontology as: "Perceived overall degree of movement of all things within the frame. This annotation type provides a scale from static to very dynamic for the rating of image-intrinsic movement."

[39] Defined in the AdA filmontology as: "The 'Field Size' is determined by the perceived size relation between a central object and the framing of a shot. This relation can be perceived as the distance towards an object of reference or how much of the centred subject in a shot and its surrounding is visible and thereby establishes the distance/proximity of the spectator to the events. Besides human bodies, reference objects can also be other figures (e.g. animals, machines). The spectrum is divided into 8 different field sizes from wide to near in accordance with Faulstich: Grundkurs Filmanalyse, 2002, Hickethier: Film- und Fernsehanalyse, 2001, Mikos: Film- und Fernsehanalyse, 2003. Additionally, there is a category for shots without a distinct reference object." For further definitions of each value see http://ada.filmontology.org/resource/2020/03/17/AnnotationType/FieldSize.html.

[40] Defined in the AdA filmontology as: "'Dialogue Text' refers to the understandable, spoken language on the audio track of a film. This usually refers to dialogue, off-commentary, but also spoken chorus. This annotation type provides a transcript of these utterances. A change of speaker or a pause marks the beginning of a new transcription unit."

[41] Defined in the AdA filmontology as: "'Music Mood' refers to the perceived emotional state conveyed in a music piece. This annotation type

provides a basic classification of the general mood that is conveyed in a coherent segment of music (either a piece or a part of it)."

Works Cited

Agt and Kutsche 2013 Agt, H. and Kutsche, R. D. "Automated construction of a large semantic network of related terms for domain-specific modeling". International Conference on Advanced Information Systems Engineering, Berlin/Heidelberg (2013), pp. 610-625.

Agt-Rickauer et al. 2018a Agt-Rickauer, H., Aubert, O., Hentschel, C., and Sack, H. "Authoring and Publishing Linked Open Film-Analytical Data", https://www.olivieraubert.net/doc/2018-ekaw-demo.pdf.

Agt-Rickauer et al. 2018b Agt-Rickauer, H., Hentschel, C., and Sack, H. "Semantic Annotation and Automated Extraction of Audio-Visual Staging Patterns in Large-Scale Empirical Film Studies". In *Proceedings of the 14th International Conference on Semantic Systems* (SEMANTICS). Vienna (2018).

Apostolidis and Mezaris 2014 Apostolidis, E. and Mezaris, V. "Fast Shot Segmentation Combining Global and Local Visual Descriptors". In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP), Florence (May 2014).

Arnold and Tilton 2019 Arnold, T., Tilton, L. "Distant viewing: Analyzing large corpora". *Digital Scholarship in the Humanities*, 34 (2019): i3-i16. fqz013, https://doi.org/10.1093/digitalsh/fqz013

Bakels 2017 Bakels, J-H. *Audiovisuelle Rhythmen. Filmmusik, Bewegungskomposition und die dynamische Affizierung des Zuschauers*. De Gruyter, Berlin/Boston (2017, english translation in prep.).

Bakels et. al in preparation Bakels, J., Grotkopp, M., Scherer, T. and Stratil, J. "Digitale Empirie? – Computergestützte Filmanalyse im Spannungsfeld von Datenmodellen und Gestalttheorie". Montage AV, 21.1 (in preparation).

Balázs 2010 Balázs, B. *Early Film Theory. Visible Man and The Spirit of Film*. Berghahn Books, Oxford (2010[1924/1930]).

Barker 2009 Barker, J. M. *The Tactile Eye. Touch and the Cinematic Experience*. University of California Press, Berkeley/Los Angeles/London (2009).

Bellour 1975 Bellour, R. "The Unattainable Text", *Screen* 16.3 (1975): 19–28.

Bellour 2000 Bellour, R. *The Analysis of Film*. Indiana University Press, Blumington, IN (2000).

Bellour 2011 Bellour, R. "Going to the Cinema with Guattari and Stern". In E. Alliez and A. Goffey (eds.), *The Guattari Effect*. Continuum, London/New York (2011): 220-234.

Berger 2019 Berger, H. "Film denkt Revolution. Zu audiovisuellen Inszenierungen politischen Wandels". *Vorwerk* 8, Berlin (2019).

Bordwell and Thompson 2013 Bordwell, D., Thompson, K. *Film Art. An Introduction*. McGraw-Hill, New York (2013).

Burdick et al. 2016 Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. and Schnapp, J.. *Digital_Humanities*. MIT Press, Cambridge (MA)/London (2016).

Bühler 1933 Bühler, K. Ausdruckstheorie. *Das System an der Geschichte aufgezeigt*. Fischer, Jena (1933).

Coppi 2002 Coppi, R. "A theoretical framework for Data Mining: The 'Informational Paradigm'", *Computational Statistics and Data Analysis*, 38.4 (2002): 501–515.

Corrigan and White 2012 Corrigan, T., White, P. *The Film Experience. An Introduction*. Bedford/St. Martin's, Boston/New York (2012).

Datta et al. 2006 Datta, R., Joshi, D., Li, J. and Wang, J. Z. "Studying aesthetics in photographic images using a computational approach". In *ECCV* (3) (2006): 288–301.

Dawes 2004 Dawes, B. Cinema Redux.http://www.brendandawes.com/projects/cinemaredux (2004).

Drucker 2016 Drucker, J. "Graphical Approaches to the Digital Humanities". In S. Schreibman, R. Siemens and J. Unsworth (eds.), *A New Companion to Digital Humanities*. Chichester: John Wiley & Sons, Chichester (2016): 238–250.

Eisenstein 1988 Eisenstein, S. "Die vierte Dimension im Film". In O. Bulgakowa (ed) *Das dynamische Quadrat. Schriften zum Film*. Hochmuth, Köln (1988): 90–108.

Eisenstein 1991 Eisenstein S. *Towards a Theory of Montage*. M. Glenny and R. Taylor (eds.). BFI Publishing, London

(1991).

Eisenstein 2006 Eisenstein, S. M. “Die Vertikalmontage (1940-1941)”. In F. Lenz (ed) *Jenseits der Einstellung. Schriften zur Filmtheorie*. Suhrkamp, Frankfurt /M. (2006): 238-300.

Ferguson 2015 Ferguson, K. L. “Volumetric Cinema”, *Transition: Jornal of Videographic Film and Moving Image Studies* 2.1 (2015).

Ferguson 2017 Ferguson, K. L. “Digital Surrealism: Visualizing Walt Disney Animation Studios”, *DHQ: Digital Humanities Quarterly*, 11.1 (2017).

Fickers 2018 Fickers, A. “Hybrid Histories. Versuch einer kritischen Standortbestimmung der Mediengeschichte”, *Annali dell'Istituto Storico Italo-Germanico in Trento (Jahrbuch des Italienisch-Deutschen Historischen Instituts in Trient)*, 44.1 (2018): 117-132.

Fickers et al. 2018 Fickers, A., Snickars, P. and Williams, M.J. “Editorial Special Issue Audiovisual Data in Digital Humanities”, *VIEW Journal of European Television History and Culture*, 7.14 (2018): 1-4. DOI: http://doi.org/10.18146/2213-0969.2018.jethc149

Fiedler 1991 Fiedler, K. “Moderner Naturalismus und künstlerische Wahrheit”. In G. Boehm, *Schriften zur Kunst I*. Wilhelm Fink, Munich (1991[1881]): 82–110.

Flückiger 2017 Flückiger, B. “A Digital Humanities Approach to Film Colors”, *The Moving Image*, 17.2 (2017): 71-93.

Gaertner 2016 Gaertner, D. “Tickets to War. Demokratie, Propaganda und Kino in den USA bis 1945”. Ph.D. thesis, Freie Universität (2016).

Greifenstein 2020 Greifenstein, S. *Tempi der Bewegung – Modi des Gefühls. Expressivität, heitere Affekte und die Screwball Comedy*. De Gruyter, Berlin/Boston (2020).

Grodal 1997 Grodal, T. K. *Moving Pictures. A New Theory of Film Genres, Feelings and Cognition*. Clarendon Press, Oxford (1997).

Grodal 2009 Grodal, T. K. *Embodied Visions. Evolution, Emotion, Culture and Film*. Oxford University Press, Oxford (2009).

Grotkopp 2017 Grotkopp, M. *Filmische Poetiken der Schuld. Die audiovisuelle Anklage der Sinne als Modalität des Gemeinschaftsempfindens*. De Gruyter, Berlin/Boston (2017).

Gruber et al. 2009 Gruber, K., Wurm, B. and Kropf, V. (eds.). *Digital Formalism: Die kalkulierten Bilder des Dziga Vertov, Maske und Kothurn*, 55 (2009).

Heftberger 2018 Heftberger, A. *Digital Humanities and Film Studies. Visualising Dziga Vertov’s Work*. Springer, Cham (2018).

Hentschel et al. 2013 Hentschel, C., Blümel I. and Sack H. “Automatic Annotation of Scientific Video Material based on Visual Concept Detection”. *Proc. 13th Int. Conf. Knowl. Manag. Knowl. Technol. - i-Know ’13*, 1-8 (2013).

Hochschild (in preparation) Hochschild, B. “Die Wahrnehmung des Anderen: Zur Begegnung mit Figuren im Verhalten von Filmen und Comics”. Ph.D. thesis, Freie Universität Berlin (in preparation).

Horst 2018 Horst, D. *Meaning-Making and Political Campaign-Advertising*. De Gruyter, Berlin/Boston (2018).

Jacobs and Fyfe 2016 Jacobs, L. and Fyfe, K. “Digital Tools For Film Analysis: Small Data”. In C. R. Acland and E. Hoyt (eds.), *The Arclight Guidebook to Media History and the Digital Humanities*. Falmer; REFRAME/Project Arclight, (2016) http://projectarclight.org/book.

Jones and Hafner 2012 Jones, R. and Hafner, C. (2012). *Understanding Digital Literacies: A Practical Introduction*. Routledge, London/ New York (2012).

Kappelhoff 2004 Kappelhoff, H. *Matrix der Gefühle. Das Kino, das Melodrama und das Theater der Empfindsamkeit*. Vorwerk 8, Berlin (2004).

Kappelhoff 2018a Kappelhoff, H. *Front Lines of Community. A Postscript to Hollywood War Cinema*. De Gruyter, Berlin/Boston (2018).

Kappelhoff 2018b Kappelhoff, H. *Kognition und Reflexion: Zur Theorie filmischen Denkens*. De Gruyter, Berlin/Boston (2018).

Kappelhoff and Bakels 2011 Kappelhoff, H. and Bakels, J.-H. “Das Zuschauergefühl. Möglichkeiten qualitativer Medienanalyse”, *Zeitschrift für Medienwissenschaft*, 5.2 (2011): 78-95.

Kappelhoff et al. 2013 Kappelhoff, H., Gaertner, D. and Pogodda, C. (eds.). *Die Mobilisierung der Sinne. Der Hollywood-Kriegsfilm zwischen Genrekino und Historie*. Vorwerk 8, Berlin (2013).

Kinkle and Toscano 2011 Kinkle, J, and Toscano, A. “Filming the Crisis: A Survey”, *Film Quarterly* 65.1 (2011): 39–51.

Krizhevsky et al. 2012 Krizhevsky, A., Sutskever, I. and Hinton, G. E. “ImageNet Classification with Deep Convolutional Neural Networks”. In *dv. Neural Inf. Process. Syst.* (2012): 1097–1105.

Lehmann 2017 Lehmann, H. *Affektpoetiken des New Hollywood. Suspense, Paranoia und Melancholie*. De Gruyter, Berlin/Boston (2017).

Li et al. 2010 Li, N., Motta, E. and Zdrahal, Z. “Evaluation of an Ontology Summarization”. (2010).

Manovich 2009 Manovich, L. *Cultural Analytics: Visualising Cultural Era of “More Media”*. Milan (2009).

Manovich 2012 Manovich, L. “How to compare one million images?” In D. Berry (ed.), *Understanding Digital Humanities*, Palgrave Macmillian, London (2012): 249-278.

Manovich 2016 Manovich, L. “The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics”, *Journal of Cultural Analytics*, May 23 (2016).

Manovich and Douglas 2009 Manovich, L., Douglas, J. “Visualizing temporal patterns in visual media: computer graphics as a Research Method”. http://softwarestudies.com/cultural_analytics/visualizing_temporal_patterns.pdf (2009).

Marks 2000 Marks, L. U. *The Skin of the Film: Intercultural Cinema, Embodiment, and the Senses*. Duke University Press, Durham/London (2000).

Mashtalir and Mikhnova 2011 Mashtalir, S. and Mikhnova, O. “Key Frame Extraction from Video: Framework and Advances”. In *Int. J. Comput. Vis. Image Process*. 4 (2014).

Meunier 2019 Meunier, J.P. “The Structures of the Film Experience: Filmic Identification”. In J. Hanich, D. Fairfax (eds.).*The Structures of the Film Experience by Jean-Pierre Meunier. Historical Assessments and Phenomenological Expansions*. Amsterdam University Press, Amsterdam (2019), pp. 32-156.

Moretti 2013 Moretti, F. *Distant Reading*. Verso, London/New York (2013).

Mueller 2012 Mueller, M. *Scalable Reading*. https://scalablereading.northwestern.edu/?page_id=22 (2012).

Müller and Kappelhoff 2018 Müller C. and Kappelhoff H. *Cinematic Metaphor. Experience – Affectivity – Temporality*. De Gruyter, Berlin/Boston (2018).

Münsterberg 2002 Münsterberg, H. “The Photoplay – A Psychological Study”. In A. Langdale (ed.), *Hugo Münsterberg on Film. The Photoplay – A Psychological Study and Other Writings*. Routledge, New York/London (2002[1916]): 45-162.

Pause and Walkowski 2019 Pause, J. and Walkowski, N.-O. “SCALABLE VIEWING – Johannes Pause und Niels-Oliver Walkowski zu digitalen Methoden und den Digital Humanities”, *Open Media Studies-Blog*. https://mediastudies.hypotheses.org/1219 (2019).

Pearlman 2009 Pearlman, K. *Cutting rhythms: Shaping the film edit*. Focal Press, New York/London (2009).

Pertersohn 2009 Petersohn, C. “Temporal video structuring for preservation and annotation of video content”. In *16th IEEE International Conference on Image Processing (ICIP)*, Cairo (2009): 93–96.

Petersohn 2008 Petersohn, C. “Logical unit and scene detection: a comparative survey”. In *Proceedings SPIE 6820, Multimedia Content Access: Algorithms and Systems II*, 682002 (2008).

Plessner 1982 Plessner, H. “Deutung des mimischen Ausdrucks. Ein Beitrag zur Lehre vom Bewußtsein des anderen Ichs”. In H. Plessner, *Gesammelte Schriften VII*. Suhrkamp, Frankfurt / M. (1982): 67–130.

Pogodda 2018 Pogodda, C. “Medientechnologie und Affekt in den Inszenierungen des Irakkrieges”. Ph.D. thesis, Freie Universität Berlin (2018).

Russakovsky et al. 2015 Russakovsky, O. Deng, J. Su, H. Krause, J. Satheesh, S. Ma, S. Huang, Z. Karpathy, A. Khosla, A. Bernstein, M. Berg, A. C. and Fei-Fei, L. “ImageNet Large Scale Visual Recognition Challenge”. In *Int. J. Comput. Vis.*, 115.3 (2015): 211–252.

Salt 1983 Salt, B. *Film Style and Technology: History and Analysis*. Starword, London (1983).

Scherer et al. 2014 Scherer, T., Greifenstein, S, and Kappelhoff, H. “Expressive Movements in Audiovisual Media. Modulating Affective Experience”. In C.Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and J. Bressem (eds), *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction*, De Gruyter Mouton, Berlin/Boston (2014): 2081–2092.

Schmitt 2020 Schmitt, C. *Wahrnehmen, fühlen, verstehen. Metaphorisieren und audiovisuelle Bilder*. De Gruyter, Berlin/Boston (2020).

Schmitt et al. 2014 Schmitt,C., Greifenstein, S. and Kappelhoff, H. “Expressive Movement and Metaphoric Meaning Making in Audio-Visual Media”. In C.Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and J. Bressem (eds), *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction*, De Gruyter Mouton, Berlin/Boston (2014): 2092–2112.

Schöch and Jannidis 2013 Schöch, C. and Jannidis, F. “Quantitative Text Analysis for Literary History – Report on a DARIAH-DE Expert Workshop”. *DARIAH-DE Working Papers*, 2 (2013).

Simmel 1959 Simmel, G. “The Aesthetic Significance of the Face”. In K. H. Wolff, *Georg Simmel, 1858–1901. A Collection of Essays, with Translations and a Bibliography*. Ohio State University Press, Columbus, OH (1959[1901]): 276–281.

Simmel 1993 Simmel, G. “Aesthetik des Porträts”. In R. Kramme and A. Cavalli, *Aufsätze und Abhandlungen, 1901–1908*. Suhrkamp, Frankfurt/M (1993[1905]): 321–332.

Sobchack 1992 Sobchack, V. *The Address of the Eye. A Phenomenology of Film Experience*. Princeton University Press, Princeton NJ (1992).

Stern 1985 Stern, D. *The Interpersonal World of the Infant*. Basic Books, New York (1985).

Stern 2010 Stern, D. *Forms of Vitality: Exploring Dynamic Experience in Psychology, the Arts, Psychotherapy, and Development*. Oxford University Press, Oxford (2010).

Stratil 2020 Stratil, J. “‘Ja es ist wieder Zeit für so ein Video’. Zur rhetorischen Situation und audiovisuellen Adressierung des Rezo-YouTube-Videos ‘Die Zerstörung der CDU’”. *mediaesthetics*, 3 (2020), https://www.mediaesthetics.org/index.php/mae/article/view/83/207.

Su et al. 2005 Su, Y., Sun, M.-T. and Hsu, V. “Global motion estimation from coarsely sampled motion vector field and the applications”. In *IEEE Transactions on Circuits and Systems for Video Technology*, 15.2 (2005).

Szeliski 2011 Szeliski, R. *Computer Vision*. Springer, London (2011).

Tan 1996 Tan, E. S. *Emotion and the Structure of Narrative Film. Film as an Emotion Machine*. Erlbaum, Mahwah, NJ (1996).

Tsivian 2009 Tsivian, Y. “Cinemetrics, part of the humanities’ cyberinfrastructure”. In M. Ross, M. Grauer and B. Freisleben (eds.), *Digital tools in media studies: Analysis and research*, transcript, Bielefeld (2009): 93–100.

Verhoeven 2016 Verhoeven, D. “Show Me the History! Big Data Goes to the Movies”. In C. R. Acland and Eric Hoyt (eds), *The Arclight Guidebook to Media History and the Digital Humanities*, London (2016).

Wundt 1880 Wundt, W. *Grundzüge der physiologischen Psychologie*. Vol 2. Wilhelm Engelmann, Leipzig (1880), 418.

Wundt 1896 Wundt, W. *Grundriss der Psychologie*. Wilhelm Engelmann, Leipzig (1896), 198.

Yosinski et al. 2014 Yosinski, J. Clune, J. Bengio,Y. and Lipson, H. “How transferable are features in deep neural networks?” In *Adv. Neural Inf. Process. Syst*. 27 (Proceedings NIPS) (2014): 1–9.

Zhang and Tomasi 1999 Zhang, T. and Tomasi, C. “Fast, robust, and consistent camera motion estimation”. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1999.