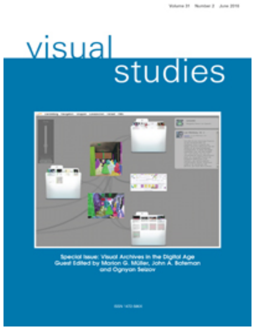


Routledge
Taylor & Francis Group

Visual Studies

ISSN: 1472-586X (Print) 1472-5878 (Online) Journal homepage: http://www.tandfonline.com/loi/rvst20

Towards next-generation visual archives: image, film and discourse

John Bateman, Chiao-I Tseng, Ognyan Seizov, Arne Jacobs, Andree Lüdtke, Marion G. Müller & Otthein Herzog

To cite this article: John Bateman, Chiao-I Tseng, Ognyan Seizov, Arne Jacobs, Andree Lüdtke, Marion G. Müller & Otthein Herzog (2016) Towards next-generation visual archives: image, film and discourse, Visual Studies, 31:2, 131-154, DOI: 10.1080/1472586X.2016.1173892

To link to this article: http://dx.doi.org/10.1080/1472586X.2016.1173892

Published online: 16 Jun 2016.

Visual Studies, 2016
Vol. 31, No. 2, 131–154, http://dx.doi.org/10.1080/1472586X.2016.1173892

Routledge
Taylor & Francis Group

Towards next-generation visual archives: image, film and discourse

JOHN BATEMAN, CHIAO-I TSENG, OGNYAN SEIZOV, ARNE JACOBS, ANDREE LÜDTKE, MARION G. MÜLLER and OTTHEIN HERZOG

The digital turn in visual studies has played a major role in the terminological overlap between 'archive', 'database' and 'corpus', and it has brought about a number of positive developments such as improved accessibility and availability. At the same time, it has also raised important questions pertaining to the materiality, searchability, annotation and analysis of the data at hand. Through a series of theoretical constructs and empirical examples, this paper illustrates the necessity and benefits of interdisciplinary dialogue when tackling the multimodal corpus annotation challenge. The meaningful interrelations between semiotic modes, the combinations between manual and (semi)automated annotation, the seamless integration of coding and annotation schemes which share common logics and the contextual embedding of the presented analyses strongly suggest multimodal document analysis in all its forms will continuously benefit from a corpus-based approach.

INTRODUCTION

In many areas of visual studies, the distinctions drawn between archives, databases and corpora have become increasingly unclear. Large-scale digitalization makes it natural for archives of visual and audiovisual material to be managed and accessed using computational techniques (cf. Schuhmacher this issue; Warnke and Dieckmann this issue). Databases of collections of visual and audiovisual material are now being maintained increasingly often, particularly where material is derived automatically or semi-automatically as in medical imaging or media studies. Furthermore, research into the formal and functional mechanisms of sophisticated combinations of visual and audial material is giving rise to evermore examples of multimodal corpora of diverse kinds and internal organisations. In this respect, the division between archives, databases and corpora is fast becoming one of disciplinary access and research methods rather than reflecting technical distinctions. But these differences in disciplinary usage can readily become misaligned with the functionalities that are actually required of the systems so described (cf. Manoff 2004, 2010).

Visual representations, such as press photography or images of paintings, for example, are still more commonly *archived* following long-established practice in art history – even though, as we will argue here, many researchers in visual communication would be better served by corpus methods. Similarly, research on speech and interaction considers itself an extension of linguistic research on texts and so employs methods and

John Bateman has been Professor of Linguistics in the English and Linguistics Departments of Bremen University since 1999, specializing in functional, computational and multimodal linguistics. His research interests include functional linguistic approaches to multimodal document design, semiotics and theories of discourse. His current interests centre on the application of functional linguistic and corpus methods to multimodal meaning making, analysing and critiquing multimodal documents of all kinds, the development of linguistically motivated ontologies and the construction of computational dialogue systems for robot–human communication.
Chiao-I Tseng is a research fellow in the Faculty of Linguistics and Literary Sciences at the University of Bremen, Germany. She is the author of the book *Cohesion in Film: Tracking Film Element* (2013, Palgrave) and several journal articles and book chapters on multimodal discourse analysis, narrative cohesion, authorship, film genre, film space and digital materiality in film.
Ognyan Seizov obtained his PhD in Communication Science at Jacobs University Bremen in 2013 (published as *Political Communication Online: Structures, Functions, and Challenges* in 2014 by Routledge). He is a Postdoctoral Research Associate at the Department of Linguistics and Literary Sciences, University of Bremen, Germany. His academic work spans political communication, multimodal website analysis, election campaigns, media reception and mixed-method research.
Arne Jacobs studied and worked at Bremen University in the TZI - Center for Computing Technologies, researching all aspects of image processing and audiovisual techniques for film and video analysis.
Andree Lüdtke studied and researched at Bremen University from 2004 until 2011, obtaining his Dr. Ing. in Artificial Intelligence, Data Mining and Databases at the TZI – Center for Computing Technologies in 2014. He is currently senior developer at neusta software development company in Bremen.
Marion G. Müller has been Associate Professor of Mass Communication at Jacobs University Bremen, Germany, since 2004. In 2009, she was founding director of the Interdisciplinary Research Center Visual Communication and Expertise (VisComX); 2007–2009 Chair of ICA's Visual Communication Studies Division. Before (2004–2007), she chaired the German Communication Association's Visual Communication Division. Her current research interests are visual political communication; visual coverage of terrorism, war and conflict; amok-school shootings and copy-cat crimes; digital prosumerism and viral video memes.
Dr. Otthein Herzog received his PhD in Computer Science from the University of Dortmund, Germany, in 1976. From 1977 to 1993, he worked for IBM Germany in various technical and managerial positions in software development, and in R&D of knowledge-based systems. From 1993 to 2009, he held the position of the chaired professor of Artificial Intelligence at the University of Bremen, Germany, where he was the founder and director of the research and technology transfer institute TZI – Center for Information and Communication Technologies from 1995 to 2009. Since 2010, he holds the Wisdom Professorship of Visual Information Technologies at Jacobs University Bremen where his research interests include the semantic analysis of images and videos, knowledge management and ICT for Smart Cities.

techniques developed in corpus linguistics although the materials managed often go well beyond those that corpus linguistics has so far evolved to deal with – that is, primarily, *linearly* organised data such as speech or text – and overlap more with materials long found in archives.

This broadening commonality both across the ways in which archive, database and corpus are used as terms and the functionalities expected or sought is not, however, always positive. As is usually the case, when terms come to overlap it is possible that useful distinctions are no longer made and opportunities for confusion follow. For example, the original sense of an archive as a collection of *material artefacts* that can then be inspected for research or other purposes is still an important and useful distinction to draw, even though an increasing number of archives no longer maintain their close link with the physical objects described (cf. Folsom 2007). Part of the problem here can be traced back to weaknesses in the theoretical foundations employed with regard to *materiality* itself – this is particularly so when electronic archives build on semiotically more traditional and 'immaterial' notions of text (cf. the valuable discussion of this point in Hayles 2003). Here multimodal corpora and, more specifically, certain now quite well-developed positions in multimodal semiotics and multimodal linguistics are beneficial to consider because these frameworks already offer a more effective placement of materiality in their treatments of signifying practices (cf. Bateman 2011).

Even worse are confusions between systems of data management and ways of employing data for particular purposes, where questions of *narrative* and *genre* have also been drawn into the mix (cf. Manovich 2001; Folsom 2007). As Hayles (2007) makes amply clear, however, very different kinds of abstractions are at issue here and any more or less metaphorical collapsing of distinctions weakens our grasp of the processes – both technical and societal – involved. Even though it may well be possible, for example, to employ structures *reminiscent* of database models for building communicative genres – as can be seen in several kinds of more experimental literature (e.g. Georges Perec) or film (e.g. Peter Greenaway) – ontologically distinct categories are involved. There is no strict dependence or equivalence holding between these levels of abstraction: after all, one can, as Hayles illustrates, just as well approach database models via more traditional narrativisation strategies as long explored in literature (cf., e.g. Genette 1980). What then is crucial here is an understanding of the importance of the role of the *interfaces* supplied to any data maintained. Awareness of this issue is also growing across disciplines (cf. McGann 2007) in order to avoid the difficulties that arise when data models (e.g. those specific to particular databases) and the manners of interaction supported in order to work with such data (interfaces) are not clearly conceptualised and related to one another appropriately.

Ontological confusions also arise in overly simple imports of notions from database design to the broader concerns of research into cultural artefacts of all kinds. Databases are sometimes characterised as 'self-describing' because there is no need (it is said) to go beyond the database in order to 'understand' what the data captured is (Hayles 2007, 1604). It is dangerous to read more into this than is actually the case – the powerful formal operations that can then be pursued to explore data have nothing to do with an 'understanding' of that data. This quickly becomes clear when the task of relating databases employing divergent database schemes for their data must be addressed and is one of the reasons motivating the current interest in incorporating formal ontologies in order to explicitly represent the *semantics* of the maintained entities and relations so as to make them available for reasoning (cf., e.g. Wache et al. 2001).

Moreover, any notion that data is maintained in obviously accessible categories is difficult to defend when the processes of search and retrieval become non-transparent to the users of those processes. It is then doubly important that search and retrieval is not only transparent but also appropriate for the uses being made of the bodies of data interrogated. Whenever archives or databases are established, and particularly when this is done in the areas of interest to the digital humanities, diverse classifications are developed so as to support search and organisation in quite different communities. These communities involve differing research interests and so the frameworks that have consequently emerged are quite diverse. And again, while necessary and beneficial, making the most of diversity requires that differences are respected and explicitly modelled.

Against this broadening background of difference and partially overlapping concerns, our primary point in this paper will be to argue that there is nevertheless now considerable potential for consolidation of the field – consolidation both of experiences and of frameworks. In fact, we consider it time for the notions of archives, databases and corpora to 'merge' to the extent that largely common mechanisms can be employed and ways of using such resources also begin to intertwine and mix, while clarifying more rigorously the distinctions that still need to be made.

To show both why and how these aims might be pursued, we will draw on results obtained in a further research project funded by the German Ministry of Education and Science (BMBF) in which we combined (1) approaches drawing on linguistically oriented multimodal research into visual and audiovisual materials, (2) traditionally more archive-oriented work into images and iconography and (3) automatic processing methods suitable for constructing large-scale databases of audiovisual materials. The organisation of the paper is then as follows. First, we introduce more formally the ideas underlying multidimensional classifications of data – ideas common to both archives and linguistic corpora and now gradually being brought to bear on database usage (section 'Archives and multilayer multimodal corpora'). We then present an example of employing the resulting techniques for research into online visual communication (section 'Example: iconographic classifications and online communication'). Next, we set out some of the benefits of combining traditional manual codings of data with automatic processing of various kinds, drawing here on the example area of audiovisual materials (section 'Combining annotation information of different logical statuses: automatic processing'). We then present a further example of employing the resulting techniques for research, drawing on the difficult issue of film genre (section 'Example: exploring the nature of genre in film'); in certain respects, this discussion will echo again the point made above that the narratives told with respect to some data cannot be confused with that data itself and the most important development to be followed now is to support *interaction* with data so as to aid research. Finally, we conclude with some brief comments on how we see the nexus archive–database–corpus developing in the future (section 'Conclusion').

ARCHIVES AND MULTILAYER MULTIMODAL CORPORA

As noted above, substantial visual archives are now being employed by a variety of distinct research communities; some of these and their development are described by the other contributions to this special issue. In all cases, a major concern is naturally how information present in archives can be searched and retrieved in order to further research. Typically, archived 'items' – whatever these may be – are classified according to more or less richly organised collections of labels. Search then proceeds by combining labels from a provided vocabulary and retrieving items so classified. The vast majority of such information is still maintained in textual form and this is also the manner of searching and indexing that is currently by and large employed. Much of the utility of such collections of material resides in the appropriateness of the labels provided for searching and the extent to which material is correctly classified according to those labels. Here issues of abstraction are important since the level of abstraction of a set of labels needs to match the kinds of research questions that are being asked. In addition, sets of labels may be organised in more or less sophisticated ways – the simplest form may be unordered collections of tags, perhaps user-generated, whereas more complex forms might rely on richly structured controlled vocabularies, potentially picking out diverse facets of the material so classified. Finally, issues of the kinds of search mechanisms available need to be raised – when those mechanisms are unclear in their working, search results may themselves potentially be skewed or non-transparent.

These problems multiply as the searching and indexing capabilities supported grow in complexity. For example, some more recent developments in archiving now decompose individual items further into more fine-grained representations, making it possible to search not only for entire works but also for motifs occurring within or across works (cf. Warnke and Dieckmann this issue). There is little doubt that this is a beneficial and necessary development as it significantly increases the value of the material maintained, but it again requires in turn differing kinds of organising labels and more sophisticated search capabilities, as well as more explicit notions of artefacts and *artefact-parts*.

It is then interesting and relevant that the forms of organisation required in such developments start showing ever closer structural relationships with the ways of structuring data long pursued within corpus linguistics. There, early notions of corpora as collections of (originally written, more recently mixed-mode) instances of language use have largely been superceded by richly annotated data that takes for granted that that data is itself internally structured in complex ways. Thus, instead of simply annotating a block of text as a string of characters, perhaps with some text structuring elements such as paragraphs as well as more global information, that is, metadata, such as text type, provenance, and so on, that block will itself have rich internal syntactic, morphological, semantic and other kinds of information explicitly represented, primarily for the purposes of linguistic research where such structures are essential. Corpus-based approaches now constitute a central methodological pillar for many areas of language research and so technical support for such approaches has exploded in recent years, ranging across storage mechanisms for large bodies of data, techniques for complex annotation and evermore sophisticated search capabilities. Many of these techniques share a common history of development with approaches to electronic text archiving and electronic publishing, with exponents such as the well-known Text Encoding Initiative (Vanhoutte and Van Den Branden 2010).

These developments have now been driven further with the advent of multimodal corpora. As linguistic research has come to consider more diverse kinds of data – moving to include spoken language, rich combinations of written and spoken language, visually manifested accompaniments to speech, such as gesture and proximity, as well as audiovisual materials more generally, such as film – many challenges have been raised and met concerning just how corpus methods can be extended. For our current purposes, it is essential to draw upon the lessons learned in designing corpora and corpus tools for such multimodal data. Since, as a rule, the greater the extent of 'enrichment' of the material selected for a corpus, the more valuable that corpus becomes, issues relating closely to questions of data transcription (cf. Flewitt et al. 2014) considered quite generally have been addressed and incorporated within a variety of annotation schemes (cf. Bateman 2014c).

Although already present to a lesser extent with purely verbal data, when moving to multimodal corpora the requirement of being able to *combine* different kinds of information reasserts itself with a new urgency: not only are distinct kinds of organisation relevant, but those organisations may draw on quite different material aspects of the data and rarely line up with each other neatly. Corpora for such data require that arbitrarily many distinct kinds of annotation be usable for common bodies of data and that each level of annotation be allowed to reflect independent perspectives on the data at hand.

The general solution adopted to this problem within corpus linguistics relies on the *Extensible Markup Language* (XML), the current recommendation of the World Wide Web Consortium for capturing structured data of all kinds. Use of XML when maintaining, accessing and visualising information allows quite diverse styles of data processing and presentation to be combined with a previously unheard of flexibility. Moreover, particularly appropriate for multimodal corpus design is the style of annotation provided by XML termed *stand-off annotation* (Thompson and McKelvie 1997). Stand-off annotation works by only allowing *indirect* reference to the data material that is to be annotated: that is, the target data material is identified by cross-references using XML attributes that identify specific positions in the data.

This approach stands in sharp contrast with *inline annotation*, where data is itself modified by incorporation of extra information. Clearly, modifying the data itself is very limited in its capabilities and quickly becomes unworkable when the extra information to be added is perhaps not mutually consistent – as in when differing or overlapping hierarchies need to be incorporated – or open-ended. In many other cases, it simply does not make sense to 'modify the data', especially when thinking of archives where material objects are being classified. Stand-off annotation is, therefore, the logical solution. The original data is untouched and all additional annotation is placed in separate files linked to the original data by means of the *cross-referencing* feature inherent to XML. Each annotation file is then free to represent its own specific kind of information and there is no problem in having many, possibly even mutually inconsistent, annotations of the same data.

The ability to add more layers of information can also be seen as allowing corpora to successively *approximate* their data, building a principled bridge between corpora and traditional archives. Moreover, this style of organisation encourages a far stronger *separation* of information and visualisation – thus bringing out the importance of the 'interface' again as emphasised above. The formal separation of layers of annotation is also, as we shall see below, important for being able to combine results more straightforwardly. In each case, analysts can simply add their own particular 'levels' of information without having to change either the originating files or any other levels of annotation. The relevance of this style of approach for constructing 'electronic' archives has now also been clearly documented, expressed by Renear (1997), for example, in terms of the necessary progression beyond the 'ordered hierarchy of content objects' view common to early electronic text archiving approaches. Multiple, and often non-aligned, hierarchies are certainly necessary.

Moving beyond this, and as we shall turn to more below, the *logical status* of any hierarchies employed may also vary. For current purposes, it will be useful to distinguish four broad categories of data enrichment: notes, transcription, coding and analysis. These all play a role in corpus-based work, may overlap and are also often equally valuable in archive-based research contexts and qualitative research more generally (cf. Denzin and Lincoln 2011).

- *Notes* are in principle the least restricted: Any information that might be useful concerning the artefact or activity under analysis might be added in any form that the analyst chooses. When this information refers to entire items, rather than to internal structure or parts of those items, it overlaps with the notion of metadata. If the form of such notes starts to be regularised in any way, then they begin to transition towards transcription, coding and analysis.
- *Transcription* is then any more systematic attempt to 'transcode' the data: For example, to represent an audio event in a written form that allows the original audio event to be re-created in more or less detail. Phonetic transcriptions are the most commonly found in linguistic corpora. In other contexts, there is an overlap with so-called 'ground data', that is, information that is considered reliable with respect to the data and which can be combined with information that is generated as hypotheses in order to check actual data properties with predicted data properties. For example, ground data for page layout might contain checked geometric measurements of layout elements and their positioning on a page. These are therefore 're-representations' of the data from particular perspectives, although there will in general always be a *reduction* in the complexity of the data with respect to the original (cf. Ochs 1979).
- *Coding* is a further step away from the data and towards more abstract analysis. Coding of data is generally carried out with respect to categories defined by a coding scheme (cf. Berger 2000; Schreier 2012). The coding scheme identifies criteria for segmentation and for the allocation of segments to categories of the coding scheme. In contrast to transcription, it is not generally possible to go from a coding to the original data in any sense: the coding scheme serves the function of grouping together segments that fall under the same classification; detail below the level of the coding scheme does not play a role beyond providing evidence for particular allocations to coding scheme categories rather than others. The units of coding may also be quite different to those of transcription – generally less fine grained.
- *Analysis* describes units identified in the data at higher levels of abstraction which typically capture how combinations of elements from the data interplay to form more complex structural configurations. For example, a corpus of spoken language could contain syntactic analyses of the units identified in the corpus. In such a case, transcription might provide a rendition of the spoken speech signal in terms of phonetic categories and lexical items; coding might attribute the lexical items to parts of speech; and analysis might run a parser over the coded lexical items to produce syntactic trees. Indeed, any of the above stages may be additionally supported by automatic processing techniques.

Finally, regardless of whether the information is made up of notes, transcriptions, codings or analyses, it must be linked in some way to the specific data which it is characterising. When the data is itself available in electronic form, then the usual XML cross-referencing mechanisms can manage this as stand-off annotation. It also becomes possible, however, to consider material archives as a further level of 'content' information – this then combines archives in a traditional sense with electronic archives that operate more like multimodal corpora. For example, one level of annotation may include geographic and other reference information that would allow actual physical artefacts to be located, other levels of annotation might then provide ground data with respect to those artefacts or transcriptions at sufficiently low levels of abstraction as to 'stand in' for the original artefacts *with respect to certain research questions*.

The approach to working with bodies of data gained and organised in this way is essentially modular – that is, distinct levels of annotation can be designed for specific tasks largely independently of one another. This also makes it straightforward to progressively accumulate analyses and annotations with respect to common data sets. It is this that then becomes crucial for the move we suggest here for productively combining database, archive and corpus research. Any kind of data can be cross-classified according to independent description schemes as motivated for addressing any particular research questions. Stand-off annotation ensures that such schemes do not become entangled with one another and remain essentially open and new schemes can always be added.

In essence, this then supports archives that allow many different kinds of access to the material they contain and matches well with the often observed 'multidimensional' nature of cultural artefacts. Folsom, for example, tries to suggest the appositeness of the rhizome model favoured by Deleuze and Guattari (1987) for cultural interpretation, that is, 'the subterranean stem that grows every which way and represents the nomadic multiplicity of identity – no central root but an intertwined web of roots' (Folsom 2007, 1573). With multilayered, modular stand-off annotations, we have an appropriate technological modelling of this notion, shorn somewhat of poetic licence and directly usable for supporting research, as we shall now see.

EXAMPLE: ICONOGRAPHIC CLASSIFICATIONS AND ONLINE COMMUNICATION

In our first example, we show how we have extended annotation schemes for visually based documents to explicitly address issues of relevance for sociopolitical analysis and interpretation. In relation to the description above, therefore, we see how sets of multilayered annotation picking out particular aspects of a body of data can be progressively refined by incorporating further, more abstract levels of description for specific research purposes.

The account presented here is drawn from the work reported in detail in Seizov (2014). One of the primary challenges met is the combination of some traditional concerns and methods of visual communication in the political domain with more corpus-oriented methods as developed within corpus linguistics and multimodal corpora design. Whereas the value of such interdisciplinary work is uncontested, there remain considerable institutional barriers preventing its effective deployment. As Herbst observes: 'the fight for legitimacy [demands] a sort of isolationism' (Herbst 2008, 605) both within and across disciplines. Hybrid fields and synthetic approaches are then rendered less attractive when it comes to resource allocation (Stanfill 2012). These institutional difficulties have prevented the development of wider and deeper cooperation between fields that share interests, methods and world views. We can then see the kind of extreme openness and modularity supported by the multilevel approach to corpus construction as one effective method for allowing different approaches *equal participation* in the research process.

The idea of taking media artefacts apart and analysing them from beginning to end is certainly not new. The search for (mass-)mediated meaning has been going strong at least since the 1940s (cf. Schramm 1997; Schreier 2012). Throughout these decades, different aspects of the texts (in the widest sense of the term) have come into analytical focus, including but not limited to structure, readability, authorship traits, discourse or propagandistic strains. All these analyses were characterised by a very narrow focus, mostly on content, which required a methodology geared towards data reduction and standardisation. While this yielded excellent results for the specific purposes of each such study, the crucial aspect of growing document complexity has remained in the background. A preoccupation with identifying what is being said superseded the concern with how it was being said – that is, despite the conviction that multimodal meaning-making relies 'on the simultaneous orchestration of diverse presentational modes, analytical methods for handling this orchestration are few and far between' (Bateman 2008, 1).

An account of document and page design focusing specifically on the contribution of layout and other visual organisational cues is set out in some detail in Bateman (2008); one result of this work is a multilevel annotation scheme, called the 'genre and multimodality (GeM)' model, which includes descriptions of visual and spatial layout in addition to considerations of linguistic 'content'. The GeM model defines the layers of description for multimodal documents shown in Table 1. The model claims that these layers are the fewest required for doing justice to page-based documents – there are certainly more, but this minimum ensures crucial components of almost any multimodal document will not be left out of the picture. The formal specification of the layers defined by the GeM model then provides the basis for the construction of

TABLE 1. The primary layers of the genre and multimodality framework for page-based artefacts.

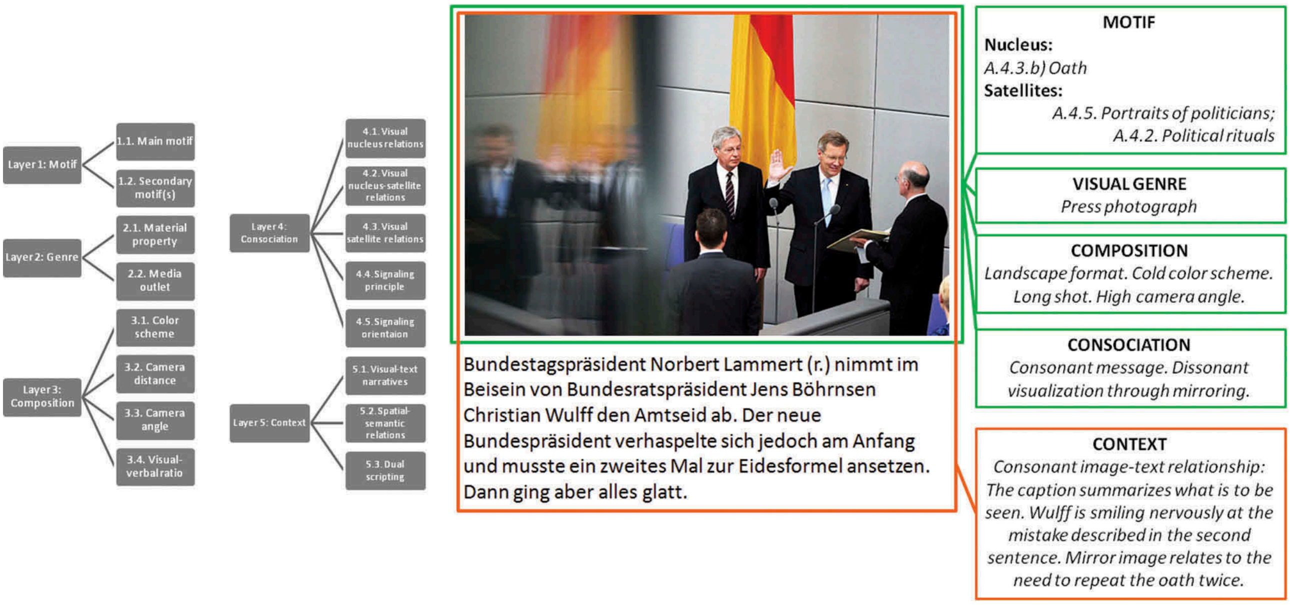
Layout structure	The nature, appearance and position of communicative elements on the page, and their hierarchical inter-relationships
Navigation structure	The ways in which the intended mode(s) of consumption of the document is/are supported: this includes all elements on a page that serve to direct or assist the reader's use of the document
Linguistic structure	The linguistic details of any verbal elements that are used to realise the layout elements of the page/ document
Content structure	The content-related structure of the information to be communicated – i.e. the propositional content or, appealing to the terms of linguistic *register theory*, the 'field' of discourse (cf. Martin 2001)
Rhetorical structure	The rhetorical relationships between content elements: i.e. how the content is 'argued', divided into main material and supporting material, and structured rhetorically
Genre structure	The individual stages or phases defined for a given genre: i.e. how the delivery of content proceeds through particular stages of activity

multimodal document corpora conforming to the accepted recommendations and standards for linguistic corpus design – in particular, the layers are each captured as XML schemes, they are independent of one another and are related as required both to each other and to the original document pages analysed by stand-off annotation. Multimodal document analysis could therefore be placed on a firmer empirical basis by selecting documents for inclusion within multimodal corpora and by 'marking up' these documents with descriptions at each of the layers proposed by the GeM model. This has supported investigation of multimodal visual genre in document design as well as design critique (cf. Bateman, Delin, and Henschel 2004; Bateman 2014a). These levels have been used to annotate several corpora of different kinds of multimodal documents (cf. Thomas 2009; Hiippala 2013, 2014), allowing some significant issues in the methodology of visual communication research and its relation to empirical methods to be addressed (Thomas 2014).

While such work has brought out generalisations concerning design and its effective deployment for communication, there are certainly many other kinds of questions that it has not addressed. One such area of concern in visual studies is that concerning political and ideological orientations in the visual messages exchanged. Precisely such a level of abstraction is however offered in the iconology-inspired account of visual motifs developed by Müller (2006, 2011). This framework, called political iconographical archive of vision (PIAV), provides an extensive description and classification hierarchy of visual styles and strategies applicable to visual communication artefacts in general. A natural question, then, is whether the approaches might be beneficially combined using the GeM model to deal with issues of increasing 'document' complexity and iconographic categories for political interpretation.

Seizov (2014) takes this question further and defines an additional 'module' of annotation specifically targeting the range of possibilities found in online communication, specifically web pages. This annotation module, called Imagery and Communication in Online Narratives (ICON), is itself multilayered, consisting of five inter-related perspectives from which the deployment of visual–textual material in online communication offerings can be characterised. The framework relies on the power of political iconology to identify and trace visual motifs and compositions and to extract meanings from them, and combines this visual approach with careful attention to the textual messages which surround the images as well as the layout and composition of the total web page space as characterised in the GeM model.

Each of the five distinct levels of ICON provides particular insights into the features of the visual and/or the image–text relationships found in the multimodal artefact under scrutiny; they are summarised graphically in Figure 1 (left) together with their immediate subcategories. Some layers operate within the boundaries of a single discipline/method, while others bridge two or more approaches in order to capture



Bundestagspräsident Norbert Lammert (r.) nimmt im Beisein von Bundesratspräsident Jens Böhrnsen Christian Wulff den Amtseid ab. Der neue Bundespräsident verhaspelte sich jedoch am Anfang und musste ein zweites Mal zur Eidesformel ansetzen. Dann ging aber alles glatt.

MOTIF
Nucleus:
A.4.3.b) Oath
Satellites:
A.4.5. Portraits of politicians;
A.4.2. Political rituals

VISUAL GENRE
Press photograph

COMPOSITION
Landscape format. Cold color scheme. Long shot. High camera angle.

CONSOCIATION
Consonant message. Dissonant visualization through mirroring.

CONTEXT
Consonant image-text relationship: The caption summarizes what is to be seen. Wulff is smiling nervously at the mistake described in the second sentence. Mirror image relates to the need to repeat the oath twice.

FIGURE 1. (Left) Five-layer organisation of the ICON classification scheme (Seizov 2014). (Right) Example of ICON analysis of a textual–visual reporting of the German politician Christian Wulff taking an oath of office (visualisation and analysis by Ognyan Seizov).

significant semantic interactions between different communication modes. Similarly, some layers deal with individual visuals distributed across a web page, while others annotate the complete document. These differences are explicitly stated in each layer's definition and are free via cross-references to draw on the detailed segmentation into layout elements provided by the GeM annotation scheme. The order of the layers is not hierarchical although they can usefully be considered in a sequence of ascending complexity and abstraction. Nevertheless, none of the layers can generate information complex enough to meet the overarching goal without the knowledge extracted from the others, so it is more accurate to speak of interconnectedness rather than supremacy of one over any of the others, precisely as emphasised in rhizome-style accounts.

The first three layers focus on the prominent visuals found on each web page and describe them in detail in terms of content (what can be seen and how it is presented; what are the presentation genre, format and style; what production values are evident in the visual). The first layer consequently picks out prominent motifs in terms of persons, actions, objects, and so on and draws directly on the detailed classification of political iconography set out in Müller's PIAV framework. The second layer identifies multimodal media such as photograph, caricature, infographic, map, and so on, as well as the 'outlet channel', for example, news, election campaign, private citizens and so on. The third layer covers compositional aspects, including technical features of design – such as colour schemes, camera distance and angles – as well as the ratio between visuals and text in order to determine the space and, hence, weight assigned to each communication mode. When the thorough descriptions of each prominent visual have been included, the fourth layer classifies the semantic relationships between visuals as well as the presence and kinds of attention guides to shed more light on the semantic structures evident in the page. Finally, the fifth layer examines the visual–verbal intersemiosis and the content organisation principles which each web page displays employing a range of potential specially defined text–image relationships.

Space precludes going through the definitions in detail and so we present here an illustrative example of a piece of communication analysed according to the ICON framework as shown in Figure 1 (right). Here we can see both visual and verbal information being combined to motivate a detailed classification of the role and manner of the contribution made by the visual elements of the layout. The classification as a whole combines the motif layer categories from PIAV, while the layout composition draws on the independent categories from GeM. Taken together, this begins to suggest something of the multiplicative effect of freely combining annotation layers. In the present example, the analysis shows that the caption is particularly useful for motivating the curious visual effect that the image reproduces since it draws attention to the fact that the oath-taker misspoke and so had to repeat the oath, an effect then echoed visually. In short, consonance and dissonance play various roles in our understanding of text–image relations and without capturing this aspect our analysis of a document remains incomplete and imprecise. ICON can therefore by itself and in combination with other annotation modules be used to enrich considerably the empirical investigation of visual–textual media.

Seizov (2014) then pursued a particular study of this kind by addressing a sample of 52 web pages sharing a political communication background and focusing on US-American political and societal topics. Each specific subsample selected for the study (NGOs, private websites, political news websites, etc.) isolated an interesting case of political communication bound by slightly different rules, expectations and overt as well as covert purposes. Nevertheless, the leading trend which recurs throughout all five ICON layers was that of clarity and coherence. More than half the web pages in the sample display visual austerity and multimodal consonance. Conflicting messages within and across semiotic modes are a rarity which deserves special attention and interpretation. The structures which support this message relay can be best characterised as straightforward and bound by design clarity. In the typical case, the visuals are produced without multiple layers of meaning, in neutral colours, from level angles; visualisation is sparse and to the point; the accompanying text is related to the visuals and often even retells their stories. Text dominates over imagery, which is another check against ambiguity. Hence, the dominant structures are multimodal, with a clear emphasis on text, and the narratives are organised according to the verbal rather than the visual flow of information and meaning. All web pages except one employ attention guides, where text dominates again, and they extend the meaning structure by emphasising and organising the concrete page's content and by providing meaningful hyperlinks to other relevant documents.

The study also showed that the major communicative function employed, for example, in the news subsample was to inform. The emphasis on clear formulations and consonant multimodal narratives speaks of a focused effort towards coherence and transparency. The visual analyses reveal mostly unimpassioned, illustrative imagery which is characteristic of professional news media, which tend to strive for objectivity. Given that such media provide less than half of the overall sample, this finding is surprising, at least at face value. Campaigns and NGOs usually aim to persuade and do not refrain from using irrational, emotional appeals, which is best done with ample visualisation. Contrary to this established practice, the overall sample contains almost no examples employing affective imagery of this kind. The main function which visuals take on is to illustrate and reinforce the verbal information, and less often to fulfil the World Wide Web's hypermodality standards by serving as attention-grabbing hyperlink illustrations. It is then intriguing to consider whether this visual 'subjugation' appears across other potential subsamples consistently.

Seizov (2014) accordingly presents more details on differences in the deployment of multimodal resources across the different media sources considered. For example, the kinds of visuals present in the campaign subsample are more varied compared to the news pages. Photographs are still the most populous category (55.32%), followed by drawings (21.28%), most of which came from Mitt Romney's campaign. Infographics and maps account for 15% altogether, and 8.51% cover miscellaneous visualisation (such as logos and symbols which do not fit any other coding category). This variety of visuals may well stem from the persuasive nature of political campaigns. Photographs are powerful transmitters of meaning, and modern visual production techniques and technologies allow for an almost unlimited variety in ways of spinning them that utilise their quality of realism for specific persuasive purposes. There is, then, evidently much more to be brought out here, although the detailed ICON classification already reveals not only generalisations, in the form of a typology of multimodal political communication types and 'visual genres', but also differences, in terms of presentational design strategies employed. As a consequence, both the findings about the overall sample and the comparisons between subsamples have the potential to generate relevant new information about political communication processes online, their structures and functions as well as the challenges they face. Versions of the ICON classifications have since been applied to other instances of online campaign communication, such as the 2012 US Presidential Election (Seizov and Müller 2015) and the 2014 Bulgarian local elections (Seizov 2015).

COMBINING ANNOTATION INFORMATION OF DIFFERENT LOGICAL STATUSES: AUTOMATIC PROCESSING

Multilayered corpus techniques certainly then provide beneficial ways of multiply indexing 'archives' of visual information: the example of the previous section applied a combined annotation scheme for characterising a body of verbal–visual online information offerings, allowing particular design strategies to be catalogued and related back to communicative purposes and contexts of production. Classifications made in such ways can then feed directly into the search for patterns, both within single levels of annotation and across distinct levels in order to ascertain reliable correlations. Traditionally, much of the work of forming such classifications for sophisticated artefacts has been done by hand – that is, bodies of coders are trained and then set to work for coding data. This was the method employed by Seizov in the previous section. Within other areas, for example, within linguistic corpora, it has in contrast become common to employ a range of *automatic* techniques for adding annotations. For larger-scale work, such automatic procedures are in fact essential since it is otherwise difficult, if not impossible, to provide sufficient quantities of appropriately classified data for more sophisticated pattern search techniques to be employed. An experimental approach compatible with automating some of the object recognition capabilities required for Seizov's analysis of static visual data was, for example, also undertaken in our overall project and is described in detail in Teichert (2011).

Automatically and semi-automatically assigned levels of description are relevant for audiovisual data in general, however. In this section, therefore, we illustrate how such levels can be achieved and also how their *combination* both with each other and with manual levels of coding can contribute further to the tools and methods we can draw on for probing the workings of visual communication. This also requires us to pick up again on the importance of providing appropriate 'interfaces' to complex annotated data or 'databases'/ 'archives' as emphasised above. To this aim, we describe an experimental interface we have developed for working with audiovisual data where automatic annotation levels are combined with more traditional, and often more abstract, manually produced annotations for exploring specific research questions.

Multiple Levels of Independent Automatic Processing Results

Manual analysis for annotation levels within audiovisual data is particularly time-consuming, and so the construction of automatic systems for audiovisual analysis is often a worthwhile goal independently of the particular research questions addressed. However, since full automation of analysis for detection of more abstract visual and audiovisual patterns is still well beyond the state of the art, methods need to be developed for investigating abstract research questions with as much support as possible from less abstract levels of coding that are approachable automatically. The key to our approach is then again to build on a framework involving an open-ended set of annotation levels just as pursued in linguistic and multimodal corpora as introduced above: in such a scheme, some annotation levels are produced by hand, others by the results of automatic processing.

Three issues then present themselves: first, how to provide automatic processing of sufficiently high quality to aid research; second, how to select automatic processing methods that are relevant for more abstract research questions; and third, how to relate the results of automatic processing to characterisations necessary for formulating more abstract research questions. We address the first issue by drawing on a set of state-of-the-art image and audio processing techniques, whose results feed directly into annotation layers expressed as stand-off annotations of the original film data. In this subsection, we characterise these layers briefly in order to give a concrete sense of the kinds of processing involved before proceeding in the sections following to their application for research.

Visual Feature Extraction

To begin, it is useful to introduce the term *visual feature*. Visual feature refers to certain properties of an image or a video that, on the one hand, can be extracted automatically (i.e. by means of an algorithm implemented in a computer program) and, on the other hand, carry useful information about the image or video in question. Such features can then be used to represent an image in later stages of an encompassing algorithm. Useful features focus on the relevant parts of an image with respect to a given task, thus filtering raw image data into a smaller set of data which is easier to handle. In terms of the levels of abstraction for additional information introduced in section 'Archives and multilayer multimodal corpora', we can see these as various types of *transcription*. This process supports in turn more guided search methods for uncovering significant patterns – that is, gradually moving into *coding* and then *analysis* proper. For current purposes and for illustration, we focus here on just two kinds of features: salient points and faces.

Salient points, also called *points of interest*, form a class of visual features which have received considerable attention in the last decade. Results in visual perception have shown that particular visual configurations are searched for during visual processing and deliver a significant portion of the information required for distinguishing images and determining their content. The most popular algorithm for the extraction of such salient point features is probably the *scale invariant feature transform (SIFT)* (Lowe 1999). Lowe's algorithm detects highly structured regions in images, which are believed to comprise more important information than unstructured image regions such as uncoloured surfaces (e.g. an empty sky or a wall in the background). The resulting feature data do not usually take less space than the original video encoded with a state-of-the-art encoder. The transformation is also not reversible – that is, information is lost.

SIFT features have certain properties that are desirable for tasks such as matching similar images and automatic motion detection, which is why we mention them here. Each image (i.e. each frame of a given film) is represented by several hundred SIFT features. These features include their spatial position and size in the image (see Figure 2, top) and a 128-dimensional vector describing the neighbourhood of the given position in the image. The similarity between two SIFT features can be computed straightforwardly via the Euclidean distance in the 128-dimensional feature space.

Automatic face detection includes the automation of the task of finding all regions in a given image which show a face and is an important topic in computer vision in its own right. In contrast to salient point detectors, which

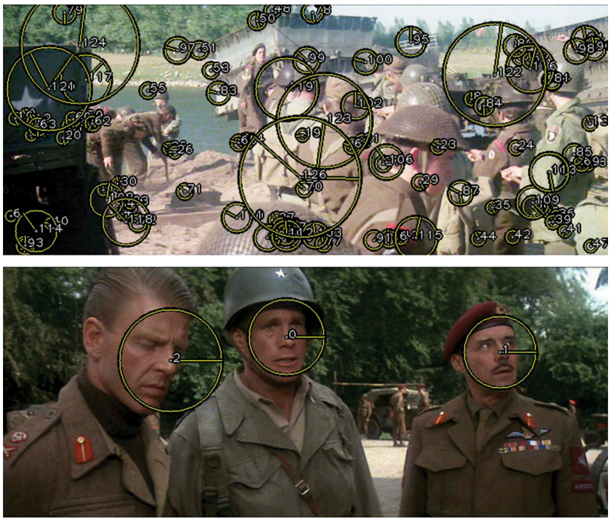


FIGURE 2. Visual features extracted from video material (sample frames taken from *A Bridge Too Far*, 1977): extracted SIFT feature points (top from 01:38:53), face detection results (bottom from 01:58:25).

extract universal image features, face detection algorithms are extremely specialised to serve a single task. This can also be motivated biologically as research suggests that humans also have parts of their visual system specialised in face detection, with even infants being able to detect faces at a very young age. The reason is, of course, that this is a vital task for every human.

Accordingly, because characters are also a vital part of almost every film genre and characters will play a central role in our illustrative example of the next section, face detection is also included as an annotation layer within our system. For this, we use a face detection algorithm proposed by Lienhart and Maydt (2002) with a corresponding model designed to detect frontal faces. As with the salient features described above, an image is represented by the set of detected face regions, that is, their spatial position and size in that image (see Figure 2, bottom). However, we do not use any further information for faces, that is, we do not distinguish between faces of different characters within this step since this would hardcode a further compounding level of possible errors. This would also reduce the possible reuse of the algorithm for detecting other patterns that may not necessarily rely on the individuation of faces. This is, therefore, indicative of our general approach of *decomposing* recognition into freely recomposable bundles of properties.

As stated above, the main function of the salient points extracted by Lowe's SIFT algorithm is to serve in later, generally more abstract, stages of automatic analysis. In our case, we use the SIFT features to determine both the similarity of successive frames of a video for the detection of *shot boundaries* and the similarity between several successive shots for *visual continuity*. We then group the shots of a video by similarity or continuity to determine sequences of shots which belong to the same *scene*. Grouping continuous shots into *scenes* is a basic step for further discrimination of scene classes for more abstract research questions – we will draw on this more in the section following.

Matching between two images is done using the approach also suggested by Lowe (1999). Each SIFT feature from the first image is compared to each feature from the second image. Given a SIFT feature from the first image, if the most similar SIFT feature from the second image (according to Euclidean distance in the 128-dimensional feature space) is significantly more similar than the second most similar SIFT feature from the second image, it is considered a valid match. We define the *absolute* similarity between two images as the number of valid matches between the SIFT features of those images.

This then helps with the detection of the boundaries between shots. Here we first compute the *absolute* similarity between a frame and its successor as well as the absolute similarity between the frame and its predecessor. We then define the *relative* similarity between two frames as the ratio of the absolute similarity between those two frames and the maximum of the absolute similarities between those two frames and their predecessors and successors. The *relative* similarity between two frames lies in the range of 0 (not similar) to 1 (very similar). To determine the actual shot boundaries, we apply a threshold on this relative similarity. If the relative similarity between two successive frames falls below the given threshold, a shot boundary is assumed. A value of 0.5 has been empirically determined with respect to our data to often be a suitable threshold, although this can vary depending on the properties of the data.

Moving on to scene boundaries, we build on our previous work on visual similarity in videos (Jacobs 2006; Jacobs, Lüdtke, and Herzog 2008) and consider the relative similarity between several shots in sequence. The relative similarity between two shots is defined as the relative similarity between the last frame of the first shot (i.e. the first regarding its temporal position in the video) and the first frame of the second shot. We again apply a threshold to the relative similarity between two shots to determine whether two shots are taken from the same camera angle. We thus get a measure of what we call visual *continuity* that allows us to group shots from the same scene.

When there is a cut to another shot from a different angle and then back to a shot from the first angle, there is a continuity between the two shots from the same angle which may be detected with our measure of visual similarity. Shots that are found similar in this way form a *bracket* around the other shots in between and are assumed together to form a continuous *scene*. Temporally overlapping brackets (i.e. groups of similar shots which alternate) are considered to belong to the same scene. This then serves to group local ABA, ABAB, and so on sequences into single units that we take as good hypotheses for individual scenes.

Audio Events and Feature Extraction

Certain audio events, that is, events that can be detected by listening only to the audio track of a video, are also automatically detected by our system using a set of computationally efficient audio features and a support vector machine classifier based on the work of Möhlmann (2007). The audio events we focus on for the examples discussed below are speech and, for reasons that will become clear, shootings, bombings and explosions. The detection of speech sequences is not to be confused with automatic speech recognition or speaker segmentation – here we are only interested in the information *that* a character is speaking, not *what* or *who* is speaking.

The continuous *scenes* that result from the algorithm described above can be further filtered to detect certain scene classes. Our system implements an algorithm for the detection of dialogue scenes based on an initial segmentation according to the visual similarity relations computed between the shots of a scene and other audiovisual features that we have computed previously (e.g. occurrence of faces and speech). Although we have tested several approaches, the one which we employ uses the visual similarity within a scene to create a similarity graph, with the shots as nodes and edges between similar shots; the use of such transition graphs is related to work such as Sidiropoulos et al. (2009). The audiovisual features, that is, the occurrence of faces, speech and/or any other acoustic properties, are associated as additional discriminating information with each shot's node in the graph. We then use a graph-matching algorithm based on spectral graph theory (Chung 1997; Wilson, Hancock, and Luo 2005) to compare each scene graph to a prototypical dialogue.

Supporting Empirical Research: An Enhanced Video Player

A summary of the combined audiovisual detectors that were explored within our system is given in Table 2. This is a considerable body of information when carried out even for single films, let alone collections of films as might be explored in an archive or corpus. It is, therefore, useful to consider how such information can be employed for further research questions since they clearly need not be particularly enlightening if considered in isolation.

This relates closely to the concerns we raised above with providing access to data exhibiting growing complexity – complexity not only of the data maintained but also of the kinds of classifications available for inspecting that data. Whereas for traditional text corpora it was often possible to take relatively simple approaches to viewing the results of corpus searches by presenting them 'directly' – that is, as collections of fragments of text meeting the search conditions – this is considerably less attractive when multimodal data is concerned. Simply replaying segments of the full, and often quite complex, multimodal artefact or behaviour according to some search criteria is unlikely to be sufficient. As a consequence, particularly in the multimodal context, appropriate tools for *interacting* with corpora take on an increased significance. Search itself needs to be seen more as explorative investigation, where partial results concerning certain aspects feed into further lines of inquiry pursuing other aspects. Here, therefore, we also begin to find considerable overlaps with the kind of *usage* made of archives, where it might not always be clear just what is being searched for and the research question develops in response to knowledge being gained along the way.

To present automatically detected technical devices and to support identification of more sophisticated event patterns, we have consequently created an evaluation prototype of an enhanced video player that, on the one hand, displays analysis results and, on the other, provides a typical interface for video display. The player interface allows for visualisation of automatically detected events as well as of manual annotation, for example, for preparation of ground-truth data for testing and the kind of research task that we illustrate in the section following. The joint presentation of layers of quite different statuses – that is, automatic analysis and manual coding – allows manual analysis to be augmented by pre-sorting data through filtering/searching for certain filmic devices that are automatically detectable and strongly indicative of more abstract complex patterns. The results of such hybrid automatic and manual audiovisual analysis can then be displayed in order to support rapid and purposive browsing of the data. This is particularly useful when dealing with a large data corpus. Presenting and

TABLE 2. Summary of event detectors explored.

Visual		Audio	
SIFT features	Lowe (2004); Stommel and Herzog (2009), Stommel (2010)	Progressive features	Möhlmann (2007)
Shot boundaries	Miene et al. (2001)	Speech	Möhlmann (2007)
Visual continuity	Jacobs, Lüdtke, and Herzog (2008)	Shouting	Möhlmann (2007)
Shot/reverse shot	Jacobs, Lüdtke, and Herzog (2008)	Screaming	Möhlmann (2007)
Face detection	Lienhart and Maydt (2002)	Shooting/bombing	Möhlmann (2007)
Face clustering	Bolme et al. (2003)	Crashes	Möhlmann (2007)
Motion detection	Brachmann et al. (2007)	Background music	Brachmann et al. (2007)
Explosions	Brachmann et al. (2007)	Explosions	Brachmann et al. (2007)
Textual inserts	Miene, Hermes, and Ioannidis (2001)	Sudden volume change	Brachmann et al. (2007)

browsing results in our enhanced video player shows directly how data-driven, functionally interpreted research of this kind can be pursued.

For use in the browser, all kinds of annotations are managed in the same way as 'detection results'. Many event detection algorithms return lists with timestamps of start and end positions and additional information such as *confidence, intensity*, and so on, while others return pairs of times and values, where the value indicates the confidence with which some property is taken to hold (e.g. at time 2.3s a face has been detected with confidence 4). The video player then allows arbitrary collections of such detection result files to be loaded in the form of csv (tab-separated) files. More traditional categories that might be provided by stand-off annotation can then be included by anchoring their opening and closing tags to times. Thus, any automatic or manual annotation result that allows representation in terms of a temporal interval or time point in the film plus some additional features can be displayed. For each analysis result, the player adds in a visual bar or graph showing the result synchronised with the video display. Navigation within the video is then as normal plus the ability to move around within the annotation segmentation bars.

The appearance of the video player with various detection results loaded is then as shown in Figure 3. Here we can see that several presentational styles are predefined for the visualisation of detection results. In the case of interval-based detectors, the identified film interval is represented as a correspondingly segmented bar in the user interface below the display of the video itself and shaded according to the confidence value; in the case of the confidence changing over time, the confidence is plotted as a continuous graph. The most complex display type shown here is that for shot similarity, which includes similarity metrics between identified frames; these are visualised in the interface by arcs joining the shots judged as similar with respect to the specified similarity threshold. More specifically, in the top-left screenshot of the figure, we can see a frame from *A Bridge Too Far* (1977) with just one annotation track loaded: the detection results for shots including the similarity judgement across shots. The individual shot csv information is shown in a scrollable list to the left of the video pane. The shot bar shown beneath the video pane depicts the arcs by which detected shot similarity is indicated. Shots linked by arcs have been classified as visually similar; we will see more use of this feature below.

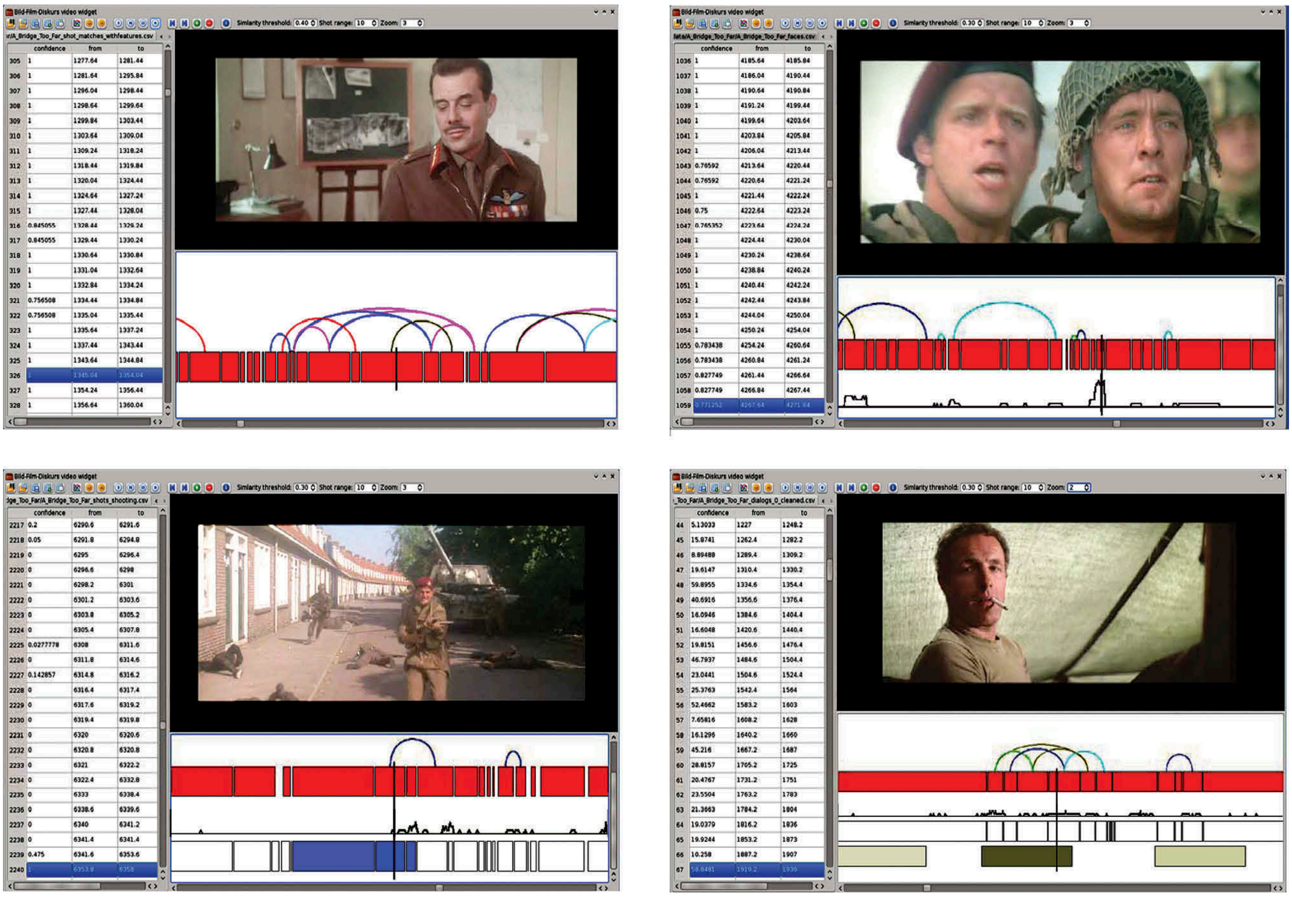


Bild-Film-Diskurs video widget

Similarity threshold: 0.40 Shot range: 10 Zoom: 3

r/A_Bridge_Too_Far_shot_matches_withfeatures.csv

	confidence	from	to
305	1	1277.64	1281.44
306	1	1281.64	1295.84
307	1	1296.04	1298.44
308	1	1298.64	1299.64
309	1	1299.84	1303.44
310	1	1303.64	1309.04
311	1	1309.24	1318.24
312	1	1318.44	1319.84
313	1	1320.04	1324.44
314	1	1324.64	1327.24
315	1	1327.44	1328.04
316	0.845055	1328.44	1329.24
317	0.845055	1329.44	1330.24
318	1	1330.64	1330.84
319	1	1331.04	1332.64
320	1	1332.84	1334.24
321	0.756508	1334.44	1334.84
322	0.756508	1335.04	1335.44
323	1	1335.64	1337.24
324	1	1337.44	1343.44
325	1	1343.64	1344.84
326	1	1345.04	1354.04
327	1	1354.24	1356.44
328	1	1356.64	1360.04

Bild-Film-Diskurs video widget

Similarity threshold: 0.30 Shot range: 10 Zoom: 3

lata/A_Bridge_Too_Far/A_Bridge_Too_Far_faces.csv

	confidence	from	to
1036	1	4185.64	4185.84
1037	1	4186.04	4190.44
1038	1	4190.64	4190.84
1039	1	4191.24	4199.44
1040	1	4199.64	4203.64
1041	1	4203.84	4205.84
1042	1	4206.04	4213.44
1043	0.76592	4213.64	4220.44
1044	0.76592	4220.64	4221.24
1045	1	4221.44	4222.24
1046	0.75	4222.64	4223.24
1047	0.765352	4223.64	4224.24
1048	1	4224.44	4230.04
1049	1	4230.24	4238.64
1050	1	4238.84	4240.24
1051	1	4240.44	4242.24
1052	1	4242.44	4243.84
1053	1	4244.04	4250.04
1054	1	4250.24	4254.04
1055	0.783438	4254.24	4260.64
1056	0.783438	4260.84	4261.24
1057	0.827749	4261.44	4266.64
1058	0.827749	4266.84	4267.44
1059	0.771252	4267.64	4271.64

Bild-Film-Diskurs video widget

Similarity threshold: 0.30 Shot range: 10 Zoom: 3

dge_Too_Far/A_Bridge_Too_Far_shots_shooting.csv

	confidence	from	to
2217	0.2	6290.6	6291.6
2218	0.05	6291.8	6294.8
2219	0	6295	6296.4
2220	0	6296.6	6298
2221	0	6298.2	6301
2222	0	6301.2	6303.6
2223	0	6303.8	6305.2
2224	0	6305.4	6307.8
2225	0.0277778	6308	6311.6
2226	0	6311.8	6314.6
2227	0.142857	6314.8	6316.2
2228	0	6316.4	6317.4
2229	0	6317.6	6319.2
2230	0	6319.4	6319.8
2231	0	6320	6320.6
2232	0	6320.8	6320.8
2233	0	6321	6322.2
2234	0	6322.4	6332.8
2235	0	6333	6338.4
2236	0	6338.6	6339.6
2237	0	6340	6341.2
2238	0	6341.4	6341.4
2239	0.475	6341.6	6353.6
2240	1	6353.8	6359

Bild-Film-Diskurs video widget

Similarity threshold: 0.30 Shot range: 10 Zoom: 2

_Too_Far/A_Bridge_Too_Far_dialogs_0_cleaned.csv

	confidence	from	to
44	5.13033	1227	1248.2
45	15.8741	1262.4	1282.2
46	8.89488	1289.4	1309.2
47	19.6147	1310.4	1330.2
48	59.8955	1334.6	1354.4
49	40.6916	1356.6	1376.4
50	16.0946	1384.6	1404.4
51	16.6048	1420.6	1440.4
52	19.8151	1456.6	1476.4
53	46.7937	1484.6	1504.4
54	23.0441	1504.6	1524.4
55	25.3763	1542.4	1564
56	52.4662	1583.2	1603
57	7.65816	1608.2	1628
58	16.1296	1640.2	1660
59	45.216	1667.2	1687
60	28.8157	1705.2	1725
61	20.4767	1731.2	1751
62	23.5504	1763.2	1783
63	21.3663	1784.2	1804
64	19.0379	1816.2	1836
65	19.9244	1853.2	1873
66	10.258	1887.2	1907
67	58.8481	1919.2	1939

FIGURE 3. Screenshots of the enhanced video player: progressively adding additional annotation tracks from left to right and top to bottom (from *A Bridge Too Far*, 1977, times: 00:22:32, 01:50:16, 01:45:56 and 00:32:04 respectively).

The second screenshot, top upper-right in the figure, then extends the information shown by adding the results of the detector for 'faces'. When faces are recognised in the image, a detection value is assigned to a time point indicating the confidence of the classification: the annotation tracks then appears here as a graph with higher values showing higher confidence. This is a useful type of depiction to use when recognition can vary continuously even within shots. The point in the particular shot shown contains two faces in close-up, and this is reflected well in the clear peak which the face detection graph indicates.

As suggested above, detection tracks can be added freely according to the research question at hand. The third screenshot, lower left, accordingly adds in the results of a 'shooting and explosion' detector: even though this is a frequent property of war films – which, for reasons that we will set out in more detail below, was one class of films that we have examined in more detail – it is nevertheless useful for separating out broad classes of events, such as battle scenes, from others. In this case, the shading of the segments indicates the confidence value: shots classified as containing gunshots or explosions with higher confidence appear darker. As can also be seen from this screenshot, confidence ratings for the faces track also start increasing just following the frame shown as the approaching soldiers come close enough for individuals to be recognised.

The results displayed thus freely combine both automatic processing of the kind we have described in this section and manual annotations with, more interestingly, *specified combinations* of automatic and manual annotations. These combinations have a number of important functions. First, whereas the ability to examine a film from the perspective of particular annotation tracks is already useful, combining results considerably enhances this functionality because of the fact that automatic detection algorithms are still far from completely reliable. This means that it is beneficial to combine sources of evidence whenever possible. We can see this played out in the last screenshot in Figure 3, shown lower right in the figure, in which a combination of annotation levels is used in order to motivate a more abstract classification of sequences into *dialogue scenes*. The reliability of recognising a dialogue scene is naturally increased by including a variety of component elements typical of dialogues: for illustration in the present case, dialogue detection might be triggered when there is a broadly alternating shot structure (as indicated in the figure by the crossing similarity arcs) and detection of faces.

This kind of track combination is suggested graphically in the close-up of the annotation bars shown in Figure 4. Whereas any of the tracks taken alone may only be a poor predictor of dialogue, their combination results in much greater prediction confidence. Systematic evaluation of the precise extent of such improvements against manually coded data still needs to be performed, although for the films examined so far the results appear quite reliable and are already of use for guiding the researcher. In fact, we predict that relying on multiple analysis results will allow many of the deficits of individual types of analysis to be counterbalanced. Since the information presented in film is in many ways redundantly coded, we can rely on this redundancy in order to extract meaningful patterns even from quite noisy individual results.

To help with defining and working with such combinations of analytical results, we also extended our enhanced video player to support the use of *detection patterns* which may freely combine classification results from other tracks. Then, again following the model of a corpus or archive use based

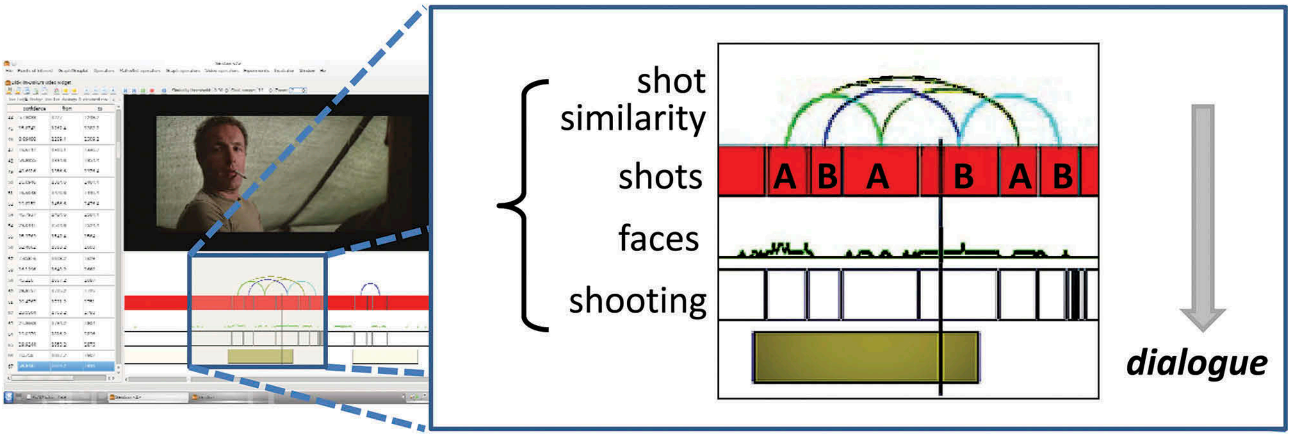


FIGURE 4. Close-up of combination of annotation tracks for detecting dialogue.

on interaction between user and data, we allowed automatically derived layers of annotation to be freely combined with manual classifications of the data so that empirically ground correlations can be sought. This is seen as one way of progressively increasing the quality and quantity of more abstract classification levels. These detection patterns are themselves expressed in a simple XML-based format and allow the combination of conditions from detection results in other tracks as well as constraints on relative temporal relations drawn from the Allen interval calculus (Allen 1984). For example, one might want to specify that intervals in which faces are recognised have to occur before intervals where speech occurs. Or, starting with basic algorithms for the extraction of visual features from a film's video track, we might show how these features can be used to determine shot boundaries and visual continuity indicating, for example, the presence of shot-reverse shot settings over a series of shots. We see a more complex usage of this facility in the example discussed in the section following.

EXAMPLE: EXPLORING THE NATURE OF GENRE IN FILM

Our second extended example will show how using combinations of automatic detection results and manual coding enables more abstract patterns to be explored in multimodal collections of data. Indeed, the use of multiple levels of annotation as evidence for further levels of description as introduced in the previous section is equally well motivated by the more abstract kinds of detection challenges that are usually required when characterising narratives. It is rarely the case, for example, that a camera angle or the sound of a gunshot alone will tell us something narratively useful; it is how such events are embedded within unfolding 'discourse contexts' that really counts (cf. Bateman 2014b). This can be employed for the evaluation of research hypotheses since a particular correlation of less and more abstract classifications may be used to specify some specific research hypothesis concerning a body of data and the generalisations that are hypothesised to hold. The more complex detection patterns that can be specified within our enhanced video player are particularly suited to this kind of exploration, as we shall now see.

Consider, for example, the perennial problem of *genre classification* for film and video. Genre expectations are commonly seen to play a central role for viewers' interpretations of films but the nature of 'genre' as such is still unclear (cf. Derrida and Ronell 1980; Altman 1984; Neale 2000). Modern accounts ranging across all media see genre as a flexible, gradual construct where varied elements of a work can more or less strongly point to prototypical generic traits. These traits may then be freely mixed and combined for aesthetic, narrative, marketing and other purposes. Under such views, genre is an intrinsically 'fuzzy' and historically variable concept. Bundles of properties at many different levels of interpretation may be combined contingently as particular prototypical genres are formed and 'named' over time. Such labels then have temporary utility as an identifier for classes of films for particular audiences (e.g. 'film noir', 'docudrama', 'splatter', 'blaxploitation', etc.). Such categories, despite a certain practical utility, remain difficult to define analytically.

Since it has become the established view that genres cannot be considered as 'natural kinds' with stable and unchanging properties and boundaries, an alternative range of approaches is called for to address the issue of genre 'bottom up', that is, by seeking systematic patterns of similarity and difference in the deployment of technical features and other significant patterns in the films themselves. This can also be seen as an orientation to corpus-driven methods since these provide a method for moving forward by refocusing analyses of genre back towards the objects of analysis *without* taking pre-existing genre labels for granted. To show one possible line of development of this kind, we suggest a new approach to genre identification and analysis that combines high-level descriptions of what we will term 'thematic configurations' with low-level combinations of visual and acoustic features. This appears to allow the establishment of groupings of films that differ systematically according to how particular thematic constructions are expressed filmically, which may in turn be used to distinguish empirically motivated 'genre' categories.

Further Narratively Relevant Filmic Annotation Layers

The ability to combine very different levels of analytic description emphasised above remains central to this approach. The manually produced levels of analysis we apply here are drawn primarily from the treatment of filmic cohesion developed by Tseng (2013b) and the corresponding framework for characterising characters' actions and interactions set out in Tseng (2013a). Each of these perspectives on film is incorporated as additional annotation levels within a multimodal corpus of film materials. Space again precludes setting out the motivations and details of these annotations here; we will, therefore, proceed primarily by example. Methodologically, our approach operates by identifying 'vertical slices' through several levels of descriptive abstraction which may then be related to one another by combining levels of annotation. The particular 'vertical slices' we focus on here all revolve around the filmic construction and use of *events*. Events as such have found increasing application in models of perception, both in natural situations and of film (cf. Zacks and Tversky 2001; Tversky, Zacks, and Martin 2008; Zacks 2010), and so it is natural to consider them here with respect to their potential for displaying genre differences also. The hypothesis that we will work towards is that differing genres may bring together and separate particular classes of events in differing ways. The events we select are driven by our theoretical considerations; their respective distributions are then pursued empirically.

To guide our exploration, we began by analysing some of the kinds of events *constructed filmically* in a body of films that might be grouped together as belonging to similar or related genres. We wanted at least some of the audiovisual properties in the films to be recognisable with a higher chance of success and so decided for investigative purposes to look at films where there were gunshots, explosions and other similarly 'loud' (both visually and audially) events. By 'filmically constructed', we refer to a regular patterning of filmic technical devices during some film in order to bring together in specific patterns of interaction certain classes of participants. Important here is that this construction is an identifiable property of a film analysed, that is, a property that can be constructed on the basis of a close 'textual' analysis of any given film. This is the style of analysis set out in detail by Tseng (2013b).

Very briefly, this analysis operates by first picking out chains of reoccurrences of particular characters and technical filmic details; these chains are called *cohesive chains*. This particular form of cohesion shows how technical devices in film identify and bring together characters, objects and settings throughout a film. Since these entities are only selected when they participate in cohesive chains, that is, they are selected repeatedly, we can state that they are constructed by the film itself as being 'salient' in some sense. This is an important means of restricting and guiding interpretation during analysis. Two distinct kinds of chains concern us for current purposes. The first tracks the filmic identities of characters, objects and settings and are termed *identity chains*; the second picks out the actions and events in which characters participate and are termed *action chains*. We use these chains to characterise how and which characters interact over the course of a film, which is then the principal analytic means we employ for defining events.

We can see this working in the following example of a sequence of analyses progressively drawing on cohesion. Figure 5 depicts a short example dialogue scene taken from Ridley Scott's *Black Hawk Down* (2001). The cohesive analysis picks out the reoccurring elements across this scene – in the present case, then, three prominent narrative elements are identified: two soldiers, here labelled A and B, and one shared setting. Each of these three entities is analysed as participating in a cohesive identity chain. These chains are summarised on the right of the figure.

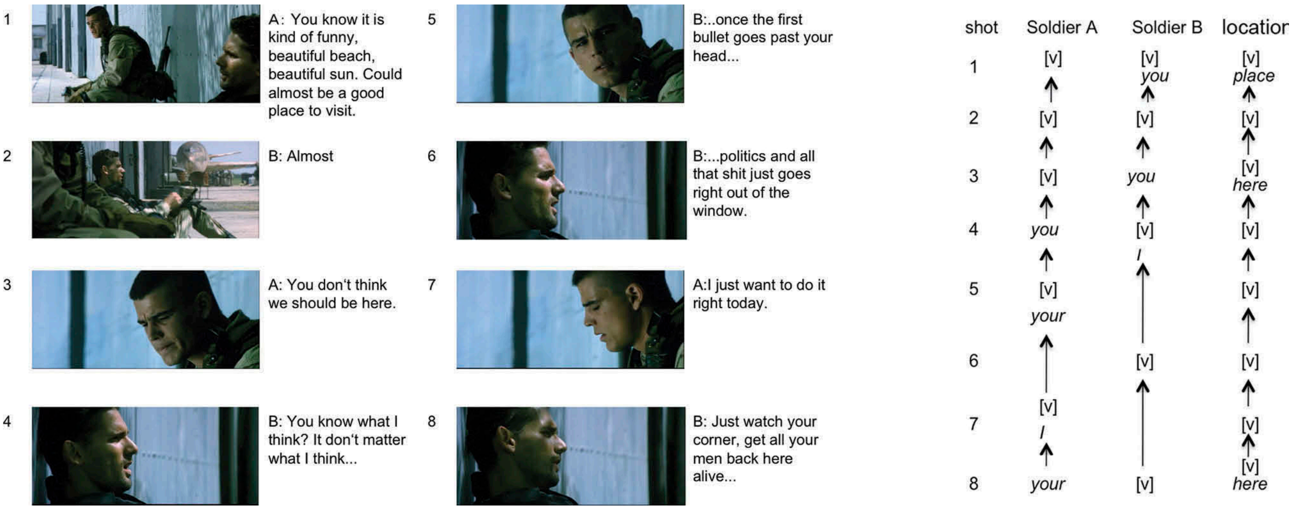

1
A: You know it is kind of funny, beautiful beach, beautiful sun. Could almost be a good place to visit.
2
B: Almost
3
A: You don't think we should be here.
4
B: You know what I think? It don't matter what I think...
5
B:..once the first bullet goes past your head...
6
B:...politics and all that shit just goes right out of the window.
7
A:I just want to do it right today.
8
B: Just watch your corner, get all your men back here alive...
shot
Soldier A
Soldier B
locatior
1
[v]
[v] *you*
[v] *place*
2
[v]
[v]
[v]
3
[v]
you
[v] *here*
4
you
[v] *I*
[v]
5
[v] *your*
[v]
6
[v]
[v]
7
[v] *I*
[v]
8
your
[v]
[v] *here*

FIGURE 5. Cohesive identity chains in a dialogue scene from *Black Hawk Down* (2001, 00:30:01–00:30:36). Key: [v] = visual figures; *italic* = spoken text.

The construction of events then operates in two steps. First, we extract action chains by selecting those portions of film where characters or figures that have been revealed to be significant by the identity chain analysis interact with each other. Again, this is a means of guaranteeing that an analysis is responding directly to the material in the film rather than any more abstract, and consequently more difficult to intersubjectively validate, interpretation. The individual interactions thus discovered are classified according to the notion of *process types* originally developed for describing language within systemic-functional grammar (Halliday and Matthiessen 2004) and subsequently extended for static (Kress and Van Leeuwen 1996) and moving images (Van Leeuwen 1996). Several distinct types of processes, with distinctively different patterns of participants and participant roles, are defined by this framework and each of these is made available as an annotation category. The action chain analysis for our example dialogue scene from *Black Hawk Down* is given in Figure 6 (top), showing a classification in terms of the process types 'mental process', 'verbal process' and 'transactional reaction'; the first chain is linguistically expressed and encompasses 'know, think, want', while the second and the third chains are realised by combinations of audiovisual cues. In general, such chains may always be constructed across a combination of expressive modalities.

The second and final step to generate the events constructed within any film under analysis is then to search for reoccurring patterns within the action chains. When such patterns are found, these are taken to correspond to classes of events that, again, are created textually by the way a film has been organised. One such pattern is shown graphically in Figure 6 (bottom), again corresponding to the portion of film in our example dialogue. This pattern brings together both the characters established in identity chains and the actions in which they participate. The nodes in the graph accordingly correspond to entire cohesive chains and not to individual audiovisual elements. This guarantees that we only consider classes of events that are constructed to have filmic 'prominence' by virtue of reoccurrences and repeated interaction across cohesive chains. The diagram in the figure shows the filmic interrelations of the two depicted characters, soldier A and soldier B. The types of interactions entered into are shown in the diagram via the process type given in the central nodes and the labels on the arcs, which give the *functional roles* of the participants according to the categories of process types employed. This is then a schematic representation of a particular class of events employed in the film analysed, one which naturally corresponds to a kind of dialogue – this is as we

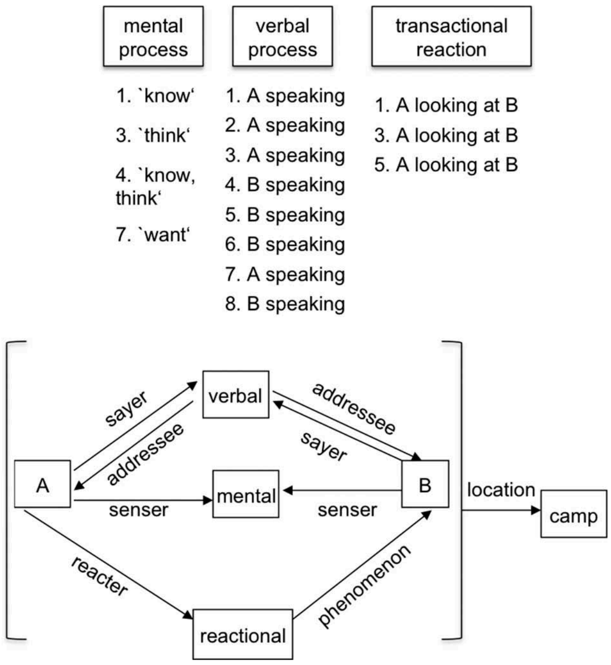


FIGURE 6. (Top) Action chains in the dialogue scene from *Black Hawk Down*. The numbers refer to the shots in the dialogue scene. (Bottom) Schematic representation of the event patterns and actant–activity relationships covering the example dialogue scene.

would have expected from an informal perusal of the material but has now also been reconstructed analytically following an intersubjectively viable method.

When this style of analysis is pursued for a film or collection of films, it is possible to isolate entire sets of filmically constructed events that reoccur. This is then one way of revealing how particular cues present within film sequences can be abstracted from to generate generic schemes of actions, roles within actions and relations between actions. It then provides a link to a more narratively relevant level of description for units of film that complements approaches to events starting from other perspectives, such as perceptual studies. Classes of events of this kind may then be considered as potential cues for questions of genre. For example, as an exploratory study we considered a small 'corpus' of six war films produced from the 1950s to 2001 with respect to the event classes they construct: *Path of Glory* (1952), *The Longest Day* (1962), *A Bridge Too Far* (1977), *Full Metal Jacket* (1987), *Saving Private Ryan* (1998) and *Black Hawk Down* (2001). In addition to generic schemes corresponding to dialogues, we could note several others that typically reoccurred, such as 'confrontation' schemes, where actual fighting occurs, and 'rescue' schemes, where some group of participants acts in order to find and retrieve another, mostly inactive (often by virtue of being wounded) participant.

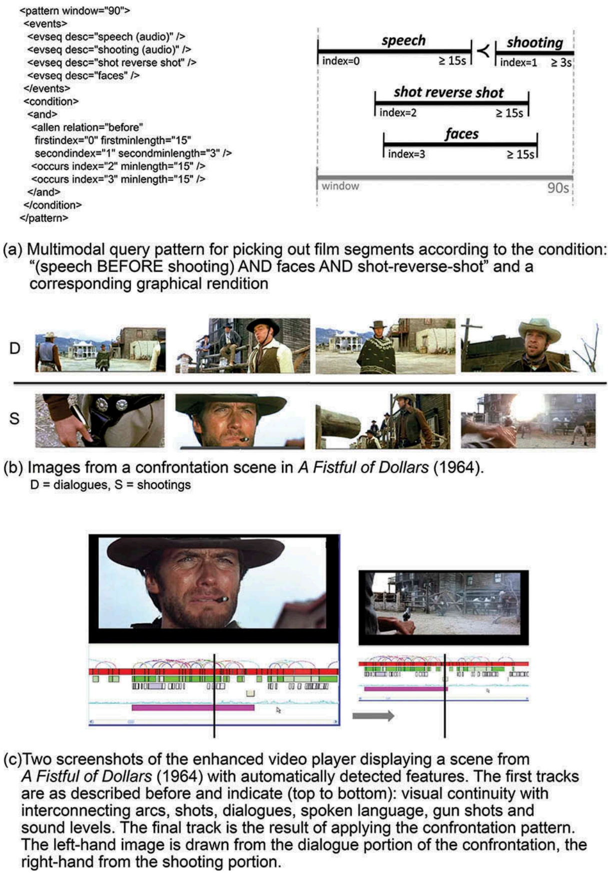
Each of the kinds of analysis discussed here, identity chains, action chains and event schema, can be seen as additional narratively relevant annotation levels that may be added into a corpus or archive of film material. Once done, this itself provides a valuable body of organised data for further exploration. Our main concern here, however, is with the use of such annotation schemes for empirical research, and in particular with the combination of manually achieved analyses such as those seen here and automatic and semi-automatic analyses in order to strengthen the empirical basis for any claims or hypotheses made. In the next subsection, therefore, we show this with respect to genre differentiation.

Event Patterns as a Tool for Genre Comparison

As a test case, we decided to examine films that are generally similar in some respects to the war films discussed above, in that they typically involve armed conflict, but which have also been discussed as a quite distinct genre in the film genre literature: that is, 'Westerns'. We considered it as a potential test for the ability of our framework to discriminate between classes of films more finely. Clearly, at the level of the existence of dialogues, confrontations and rescue events, there would not appear to be any great difference between the two classes of films: such events may occur in Westerns just as they occur in war films. In addition, in order to see whether any hypothesised patterns contribute to genre discrimination at all, it is necessary to apply the method to a larger sample of data. The descriptions given in the previous subsection are sufficiently abstract that they cannot at the present time be simply extracted from the data. Therefore, we have explored how to provide active support for using such abstract levels by combining levels of automatic and manual coding components as argued in previous sections in order to allow more ready recognition of distinctive patterns across potentially different filmic genres.

One of the 'generically' most typical kinds of events that we can observe in Westerns is that of the 'shoot-out': the class of events, often arriving at some point of narrative climax or resolution, where there is a gunfight between some of the main protagonists of the film (cf. Wright 1975). We consequently considered how we might recognise such events with the support of the automatic and semi-automatic analyses reported above. To aid empirical investigation, we encoded the result as an event detection pattern as supported by our prototype enhanced video player introduced in section 'Supporting empirical research: an enhanced video player'. This is shown in both XML and a corresponding informal graphical style in Figure 7(a). As suggested above, this event detection pattern can itself be considered as both a result of analysis and as an empirical hypothesis: it is a result of analysis in the sense that our manual analyses of films suggest that these are the properties that hold and it is an empirical hypothesis in the sense that when we apply this pattern to concrete films we can measure its success in picking out what we would want to label as 'shoot-outs'.

The first portion of the detection pattern as shown in the figure (<events> … </events>) defines the annotation levels, or tracks, that are to be examined. In the present case, we see four detection event sequences (<evseq>) being selected: one identified by the 'speech' detector, one by the 'shooting' detector, one by the itself more complex 'shot reverse shot' detector described above and one by the 'face' recognition detector. The information under the <conditions> … </conditions> part of the pattern then states precisely when the overall pattern is to be taken as 'firing' or applying. Two kinds of conditions are given: simple occurrence, that is, that one of the used event detectors should have returned positive results, and temporal ordering of positive results



```
<pattern window="90">
 <events>
  <evseq desc="speech (audio)" />
  <evseq desc="shooting (audio)" />
  <evseq desc="shot reverse shot" />
  <evseq desc="faces" />
 </events>
 <condition>
  <and>
   <allen relation="before"
    firstindex="0" firstminlength="15"
    secondindex="1" secondminlength="3" />
   <occurs index="2" minlength="15" />
   <occurs index="3" minlength="15" />
  </and>
 </condition>
</pattern>
```

(a) Multimodal query pattern for picking out film segments according to the condition: “(speech BEFORE shooting) AND faces AND shot-reverse-shot” and a corresponding graphical rendition

(b) Images from a confrontation scene in *A Fistful of Dollars* (1964).
D = dialogues, S = shootings

(c)Two screenshots of the enhanced video player displaying a scene from *A Fistful of Dollars* (1964) with automatically detected features. The first tracks are as described before and indicate (top to bottom): visual continuity with interconnecting arcs, shots, dialogues, spoken language, gun shots and sound levels. The final track is the result of applying the confrontation pattern. The left-hand image is drawn from the dialogue portion of the confrontation, the right-hand from the shooting portion.

FIGURE 7. (a) Multimodal query pattern for picking out film segments according to the condition: ‘(speech BEFORE shooting) AND faces AND shot-reverse-shot’ and a corresponding graphical rendition. (b) Images from a confrontation scene in *A Fistful of Dollars* (1964). D = dialogues; S = shootings. (c) Two screenshots of the enhanced video player displaying a scene from *A Fistful of Dollars* (1964) with automatically detected features. The first tracks are as described before and indicate (from top to bottom): *visual continuity* shown with interconnecting arcs, shots, dialogues, spoken language, gunshots and sound levels. The final track shows the result of applying the confrontation pattern. The left-hand image is drawn from the dialogue portion of the confrontation, and the right-hand image from the shooting portion.

among the event detectors. The event detectors referred to are identified by the ‘index’ attributes in the individual conditions: these refer to the specified annotation levels in the <events> list, beginning numbering from zero. Thus, the first condition specifies a necessary temporal relation between the speech recognition track and the shooting track; the remaining conditions simply state that there should also be a simultaneous shot/reverse shot sequence and recognised faces. Minimum lengths of successful event detections are also given as well as an overall time window within which conditions should be checked (the first line of the pattern). These relationships are summarised in the graphic on the right of the figure: temporal intervals are set out as horizontal bars identified according to the annotation tracks that define them. When this detection pattern is applied to a film, the results appear as additional analysis bars within our enhanced video player that can be viewed and combined with other bars in the usual way. This offers a method for isolating film segments exhibiting just those conditions that we wish to explore.

As an example of the application of the detection pattern, let us consider an in many ways prototypical film among those we have analysed: Sergio Leone’s *A Fistful of Dollars* (1964). Figure 7(b) shows in two rows (left to right and top to bottom) representative images from a segment where the event detection pattern matches. This segment turns out to be, as would be expected, a typical duel-like scene, depicting confrontation and shoot-out between the main character played by Clint Eastwood and some opponents. The scene starts with Clint Eastwood’s character walking into a village and approaching his enemies at a house; a brief conversation ensues, followed by an exchange of shots. The images in the upper row of the figure correspond to the first target interval of the detection pattern: that is, there is a shot/reverse shot sequence, with faces recognised because of two or three salient identifiable characters in close or medium-close shots and spoken language co-present because of the accompanying dialogue. This is then also stylistically similar to the dialogue scene from *Black Hawk Down* above, which would also match this particular component of the detection pattern. The lower row of the figure shows the portion of the segment matching the second interval of the detection pattern, in which spoken language is replaced by shots, explosions, and so on. In the film, these follow immediately upon one another within the overall shot/reverse shot structure and so satisfy the temporal constraint given in the pattern.

Since the event detection pattern as given above in Figure 7(a) matches, loading these results into the enhanced video player gives rise to a further analysis bar as shown in Figure 7(c). More specifically, the automatically segmented video shots are displayed in the first track (as red bars in colour versions of the figure); visual continuity based on visual similarity of start and end frames of successive video shots is indicated by interconnecting arcs as explained above. A sequence of visually continuous shots forms a *scene*. Video segments with automatically detected dialogues are displayed in the second track (as green bars). Video segments with detected shootings resulting from audio event recognition (section ‘Audio events and feature extraction’) are visualised in the third track (as purple bars). It is then precisely this *combination* of features that is picked up when attempting to find matches for the defined event detection pattern, giving rise to the lowest analysis bar.

Now, finally returning to the question of genre and possible patterns that might be of significance for distinguishing genres, when we apply the defined pattern to Westerns, we find shoot-out scenes of this kind. However, when we apply the same pattern to the war films in our sample, we find almost no cases of the pattern matching. This comparison between the relative distribution of this class of events in the two collections of films may thus reflect a qualitative difference between narrative configurations in war and in Western films. While in war films transactional actions such as shooting and combat actions are realised by salient characters (soldiers) against general enemies, transactional actions in the Western film are interactions between salient characters which still preserve or are even contiguous with their *verbal* interactions. In other words, the physically more violent actions in battlefields are performed in war films against unidentifiable people in the long shots, while in the Western film these violent confrontations and shootings are realised with identifiable characters portrayed in closer shots and often already engaging in a verbal interaction.

This also provides empirical material that could be used to support a differentiation in the narrative function of dialogue scenes across the two genres. The event analysis above for *Black Hawk Down* showed a quite common occurrence for war films where the less dynamic dialogue scenes between main characters take place in settings different from battlefields or, at least, in situations where there is not active fighting. In war films, verbal interactions between the main characters are therefore often not directly related to physically violent actions of shooting or combat. In rather marked contrast to this, the static dialogue between cowboys in the Western film often transforms seamlessly, and sometimes quite abruptly, into an exchange of gunshots by those very participants who were previously interacting verbally. Indeed, the static dialogue between the main characters in the Western film functions as one reason/motivation that triggers the physical violence. Figure 8 shows this distinctive patterning across the two genres graphically, building on the analysis tracks used so far.

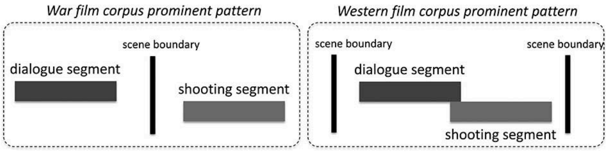


FIGURE 8. Comparison of the manipulation of scenes and events across war and Western films.

These differences may then be suggested as potential sensitive indicators distinguishing collections of films in ways that align usefully with genre classifications and which, furthermore, allow us to explore in more detail the often proposed, though rarely successfully demonstrated, connection between genres and societal meanings. This may be employed in a further argument suggesting how the general social construction of violence is constituted differently in the two classes of films we have examined. Differences between how events are constructed in our collection of films may turn out to be indicative of genre 'boundaries'. The definition and realisation of violence in moving images is itself of course a complex and widely discussed issue (cf. Prince 2003, 2009; Hartmann and Vorderer 2010). Here our analytical model has been used to suggest how the different constructions and narrative embeddings of violence may offer a potential 'dimension' of genre classification that reliably distinguishes prototypical Western and war film genres 'bottom-up'.

Within this framework, genres are consequently characterised more in terms of their 'styles of meaning', which can in turn be defined as *rhetorical strategies* (cf. Lemke 1999; Bateman 2008, 2014a), rather than in terms of superficial narrative descriptions or arbitrary labels. Hybridisation is then itself naturalised as the default state of affairs: the films grouped according to some particular genre label may then naturally change over time as styles of meaning presentation are bundled and re-bundled through use. This offers an empirically based model that accepts as a fundamental premise Steve Neal's proposal that the 'conventions of a genre are always *in* play rather than being simply *re*played' (Neale 1990, 171). This then also begins to suggest quite concretely how genre comparison might bring to light social/cultural meanings on higher, more abstract levels of interpretation.

CONCLUSION

The interdisciplinary cooperation results set out in this paper circumscribe a growing theoretical and methodological overlap between the notions of archive, corpus and database. They also point in several directions for crucial future development with regard to those terms. For the purposes of multimodal research, archives will need to move towards supporting many of the functionalities supplied by multilevel multimodal corpora. Simultaneously, corpus-based approaches will need to move away from simple notions of a corpus as a 'collection of transcribed data' where the transcriptions more or less stand in for the phenomena to be studied and instead increasingly become repositories of data seen from a variety of perspectives.

This transition constitutes a major challenge for both areas. The levels of abstraction included will need to range from low-level technical features (e.g. for spoken language: acoustic properties; for images, photographic renditions, colour palettes, etc.; for film: optical flow, colour balance, edge detection, cut detection, etc.), through transcriptions of selected perspectives on the data (e.g. for language: phonetics and intonation) and the results of experimental studies (e.g. for images or film: eye-tracking data), to more abstract analyses (e.g. for visuals: classifications of iconography, motifs, etc.; and for documents: interactions between rhetorical relations and distance and position relations between layout units, types of typographical realisations and eye-tracking predictions). Examples of correspondences across levels being drawn with behavioural measurements ascertained empirically can be found with respect to eye-tracking in Müller, Kappas, and Olk (2012), Hiippala (2012) and Kluss et al. (2016). These shed light on the vast richness of data to be drawn from multimodal corpora while at the same time illustrating the need to open up the theoretical and methodological playing field in order to achieve insights unattainable for previous mono-disciplinary approaches.

Supporting access to combinations of information (visual–verbal, visual–auditory, layout, etc.) and the search for meaningful patterns is itself complex and new methods and techniques of visualisation will be crucial (cf. Caldwell and Zappavigna 2011; Manovich 2012; O'Halloran, E and Tan 2014). Properly annotated data may then support the search for generalisations by allowing examination of potential *correlations* across the various levels of descriptions that corpora provide. The keyword here is *properly*: reproducibility and reliability remain the main issues ahead for novel corpus-based multimodal approaches, and method development and testing are of paramount importance for interdisciplinary cooperation to flourish. Establishing good practices capable of homing in on the complex interdependencies typically found in multimodal media artefacts today represents a crucial step towards understanding how visual communication and, indeed, multimodal meaning-making in general, operates. For larger-scale corpus and archive work, it will be essential to pursue an increasing utilisation and development of automatic methods in appropriate combination with manual approaches. The current paper has presented several examples of academic research already moving in this direction. Simplified and sharpened searchability in combination with improved usability and access is key to the successful integration of such automatic and manual annotation efforts. As pointed out in Thomas (2014), only large-scale corpus research where low-level annotation is automated, thereby postponing human interpretation for later phases of analysis, will be able to produce the unique brand of reliable and reproducible multimodal content or discourse analysis that the current paper has argued to be necessary.

ACKNOWLEDGEMENTS

The work reported in this paper was undertaken in the cooperative project 'Visuals-Film-Discourse' (Bild-Film-Diskurs: Bateman et al. 2012) funded by the German Ministry for Education and Research (BMBF) at the University of Bremen and Jacobs University Bremen with Co-Principal Investigators John Bateman (Linguistics, Bremen University), Marion G. Müller (Visual Communication, Jacobs University Bremen), Rainer Malaka (Digital Media and Informatics, Bremen University) and Otthein Herzog (TZI, Bremen University and Jacobs University Bremen). The enhanced video player prototype was developed by Arne Jacobs and Andree Lüdtke of the TZI, Bremen, as part of the project.

FUNDING

The work reported in this paper was undertaken in the cooperative project 'Visuals-Film-Discourse' (Bild-Film-Diskurs: Bateman et al. 2012) funded by the German Ministry for Education and Research (BMBF) at the University of Bremen and Jacobs University Bremen with Co-Principal Investigators John Bateman (Linguistics, Bremen University), Marion G. Müller (Visual Communication, Jacobs University Bremen), Rainer Malaka (Digital Media and Informatics, Bremen University) and Otthein Herzog (TZI, Bremen University and Jacobs University Bremen).

REFERENCES

Allen, J. F. 1984. "Towards a General Theory of Action and Time." *Artificial Intelligence* 23: 123–154. doi:10.1016/0004-3702(84)90008-0.

Altman, R. 1984. "A Semantic/Syntactic Approach to Film Genre." *Cinema Journal* 23 (3): 6–18. doi:10.2307/1225093.

Bateman, J. A. 2008. *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. London: Palgrave Macmillan.

Bateman, J. A. 2011. "The Decomposability of Semiotic Modes." In *Multimodal Studies: Multiple Approaches and Domains*, edited by K. L. O'Halloran and B. A. Smith, 17–38. London: Routledge.

Bateman, J. A. 2014a. "Genre in the Age of Multimodality: Some Conceptual Refinements for Practical Analysis." In *Evolution in Genres: Emergence, Variation, Multimodality*, edited by P. Evangelisti-Allori, V. K. Bhatia, and J. A. Bateman. Frankfurt am Main: Peter Lang.

Bateman, J. A. 2014b. "Looking for What Counts in Film Analysis: A Programme of Empirical Research." In *Multimodal Communication*, edited by D. Machin, 301–330. Berlin: Mouton de Gruyter.

Bateman, J. A. 2014c. "Using Multimodal Corpora for Empirical Research." In *The Routledge Handbook of Multimodal Analysis*, edited by C. Jewitt, 238–252. London: Routledge.

Bateman, J. A., J. L. Delin, and R. Henschel. 2004. "Multimodality and Empiricism: Preparing for a Corpus-Based Approach to the Study of Multimodal Meaning-Making." In *Perspectives on Multimodality*, edited by E. Ventola, C. Charles, and M. Kaltenbacher, 65–87. Amsterdam: John Benjamins.

Bateman, J. A., M. G. Müller, R. Malaka, and O. Herzog. 2012. "Image - Film - Discourse /Bild - Film – Diskurs. " TZI-Bericht Nr. 68. Bremen: TZI, Universität Bremen.

Berger, A. A. 2000. *Media and Communication Research Methods: An Introduction to Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage Publications.

Bolme, D. S., J. R. Beveridge, M. Teixeira, and B. A. Draper. 2003. "The CSU Face Identification Evaluation System: Its Purpose, Features, and Structure." In *Computer Vision Systems*, edited by J. Crowley, J. Piater, M. Vincze, and L. Paletta, 304–313. Berlin: Springer.

Brachmann, C., H. I. Chunpir, S. Gennies, B. Haller, T. Hermes, O. Herzog, A. Jacobs, et al. 2007. "Automatic Generation of Movie Trailers Using Ontologies." *IMAGE - Journal of Interdisciplinary Image Science* 5: 117–139.

Caldwell, D., and M. Zappavigna. 2011. "Visualizing Multimodal Patterning." In *Semiotic Margins: Reclaiming Meaning*, edited by S. Dreyfus, S. Hood, and M. Stenglin, 229–242. London: Continuum.

Chung, F. R. K. 1997. *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*. Providence, RI: American Mathematical Society.

Deleuze, G., and F. Guattari. 1987. *A Thousand Plateaus: Capitalism and Schizophrenia*. Minneapolis: University of Minnesota.

Denzin, N. K., and Y. S. Lincoln, eds. 2011. *The SAGE Handbook of Qualitative Research*. Thousand Oaks, CA: Sage Publications.

Derrida, J., and A. Ronell. 1980. "The Law of Genre." *Critical Inquiry* 7 (1): 55–81. doi:10.1086/448088.

Flewitt, R., R. Hampel, M. Hauck, and L. Lancaster. 2014. "What are Multimodal Data Collection and Transcription?" In *The Routledge Handbook of Multimodal Analysis*, edited by C. Jewitt, 44–59. London: Routledge.

Folsom, E. 2007. "Database as Genre: The Epic Transformation of Archives." *PMLA: Publications of the Modern Language Association of America* 122 (5): 1571–1579. doi:10.1632/pmla.2007.122.issue-5.

Genette, G. 1980. *Narrative Discourse*. Ithaca, NY: Cornell University Press.

Halliday, M. A. K., and C. M. Matthiessen. 2004. *An Introduction to Functional Grammar*. London: Edward Arnold.

Hartmann, T., and P. Vorderer. 2010. "It's Okay to Shoot a Character: Moral Disengagement in Violent Video Games." *Journal of Communication* 60: 94–119. doi:10.1111/jcom.2010.60.issue-1.

Hayles, N. K. 2003. "Translating Media: Why We Should Rethink Textuality." *The Yale Journal of Criticism* 16 (2): 263–290. doi:10.1353/yale.2003.0018.

Hayles, N. K. 2007. "Narrative and Database: Natural Symbionts." *PMLA: Publications of the Modern Language Association of America* 122 (5): 1603–1608.

Herbst, S. 2008. "Disciplines, Intersections, and the Future of Communication Research." *Journal of Communication* 58 (4): 603–614. doi:10.1111/jcom.2008.58.issue-4.

Hiippala, T. 2012. "Reading Paths and Visual Perception in Multimodal Research, Psychology and Brain Sciences." *Journal of Pragmatics* 44 (3): 315–327. doi:10.1016/j.pragma.2011.12.008.

Hiippala, T. 2013. "Modelling the Structure of a Multimodal Artifact." PhD diss., University of Helsinki.

Hiippala, T. 2014. "Genre Analysis." In *Interactions, Images and Texts: A Reader in Multimodality*, eds S. Norris and C. D. Maier, 111–123. Berlin: Mouton de Gruyter.

Jacobs, A. 2006. "Using Self-Similarity Matrices for Structure Mining on News Video." In *Proceedings of the 4th Hellenic Conference on Artificial Intelligence (SETN'06)*. http://www.tzi.de/~jarne/jacobs06.pdf

Jacobs, A., A. Lüdtke, and O. Herzog. 2008. "Inter-Video Similarity for Video Parsing." In *Intelligent Information Processing IV'*, edited by Z. Shi, E. Mercier-Laurent, and D. Leake, 174–181. Boston: Springer.

Kluss, T., J. A. Bateman, H.-P. Preußer, and K. Schill. 2016. "Exploring the Role of Narrative Contextualization in Film Interpretation: Issues and Challenges for Eye-Tracking Methodology." In *Making Sense of Cinema: Empirical Studies into Film Spectators and Spectatorship*, edited by C. D. Reinhard and C. J. Olson, 257–284. New York: Bloomsbury Academic.

Kress, G., and T. Van Leeuwen. 1996. *Reading Images: The Grammar of Visual Design*. London: Routledge.

Lemke, J. L. 1999. "Typology, Topology, Topography: Genre Semantics." MS University of Michigan. Accessed 30 April 2016. http://www-personal.umich.edu/~jaylemke/papers/Genre-topology-revised.htm

Lienhart, R., and J. Maydt. 2002. "An Extended Set of Haar-Like Features for Rapid Object Detection." Proceedings of ICIP 2002, Rochester, NY, September 22–25, 900–903. IEEE.

Lowe, D. G. 1999. "Object Recognition from Local Scale-Invariant Features." Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV), Kerkyra, September 20–27, 1150–1157. IEEE.

Lowe, D. G. 2004. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision* 60 (2): 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Manoff, M. 2004. "Theories of the Archive from Across the Disciplines." *Portal: Libraries and the Academy* 4 (1): 9–25. doi:10.1353/pla.2004.0015.

Manoff, M. 2010. "Archive and Database as Metaphor: Theorizing the Historical Record." *Portal: Libraries and the Academy* 10 (4): 385–398. doi:10.1353/pla.2010.0005.

Manovich, L. 2001. *The Language of the New Media*. Cambridge, MA: MIT Press.

Manovich, L. 2012. "How to Compare One Million Images." In *Understanding Digital Humanities*, edited by D. M. Berry, 249–278. London: Palgrave Macmillan.

Martin, J. R. 2001. "Language, Register and Genre." In *Analysing English in A Global Context: A Reader*, edited by A. Burns and C. Coffin, 149–166. London: Routledge.

McGann, J. 2007. "Database, Interface, and Archival Fever." *PMLA: Publications of the Modern Language Association of America* 122 (5): 1588–1592.

Miene, A., A. Dammeyer, T. Hermes, and O. Herzog. 2001. "Advanced and Adapted Shot Boundary Detection." In *Proceedings of ECDL WS Generalized Documents*, edited by D. W. Fellner, N. Fuhr, and I. Witten, 39–43. Darmstadt: ECDL.

Miene, A., T. Hermes, and G. T. Ioannidis. 2001. "Extracting Textual Inserts from Digital Videos." Proceedings of the Sixth International Conference on Document Analysis and Recognition (IDCAR'01), Seattle, WA, September 10–13, 1079–1083.

Möhlmann, D. 2007. "Sound Detection and Recognition Using Progressive Feature Extraction." Master's diss., Informatics and TZI, University of Bremen.

Müller, M. G. 2006. "Die Ikonographie Des Politischen Händedrucks." In *Freundschaft, Motive Und Bedeutungen*, edited by S. Appuhn-Radthke and E. Wipfler, 205–215. Munich: Zentralinstitut für Kunstgeschichte.

Müller, M. G. 2011. "Iconography and Iconology as a Visual Method and Approach." In *Handbook of Visual Research Methods*, edited by E. Margolis and L. Pauwels, 283–297. London: Sage.

Müller, M. G., A. Kappas, and B. Olk. 2012. "Perceiving Press Photography: A New Integrative Model, Combining Iconology with Psychophysiological and Eye-Tracking Methods." *Visual Communication* 11 (3): 307–328. doi:10.1177/1470357212446410.

Neale, S. 1990. "Questions of Genre." *Screen* 31 (1): 45–66. doi:10.1093/screen/31.1.45.

Neale, S. 2000. *Genre and Hollywood*. London: Routledge.

Ochs, E. 1979. "Social Foundations of Language." In *Discourse Processes: Advances in Research and Theory. Volume 2: New Directions in Discourse Processing*, edited by R. O. Freedle, 207–221. Norwood, NJ: Ablex.

O'Halloran, K. L., M. K. L. E. and S. Tan. 2014. "Multimodal Analytics: Software and Visualization Techniques for Analyzing and Interpreting Multimodal Data." In *The Routledge Handbook of Multimodal Analysis*, edited by C. Jewitt, 386–396. London: Routledge.

Prince, S. 2003. *Classical Film Violence: Designing and Regulating Brutality in Hollywood Cinema, 1930-1968*. New Brunswick, NJ: Rutgers University Press.

Prince, S. 2009. "Violence." In *The Routledge Companion to Philosophy and Film*, edited by P. Livingston and C. Plantinga, 279–288. London: Routledge.

Renear, A. 1997. "Out of Praxis: Three (Meta)Theories of Textuality." In *Electronic Text: Investigations in Method and Theory*, edited by K. Sutherland, 107–126. Oxford: Clarendon Press.

Schramm, W. 1997. *The Beginnings of Communication Study in America: A Personal Memoir*. Thousand Oaks, CA: Sage.

Schreier, M. 2012. *Qualitative Content Analysis in Practice*. London: Sage.

Seizov, O. 2014. *Political Communication Online: Structures, Functions, and Challenges*. New York: Routledge.

Seizov, O. 2015. "Communicative and Persuasive Strategies in the Bulgarian Parliamentary Elections 2014: A Multimodal Analysis." *International Journal of E-Politics* 6 (2): 43–68. doi:10.4018/IJEP.

Seizov, O., and M. G. Müller. 2015. "Multimodal Online Strategies in the US Presidential Election 2012: A Content Analysis of Barack Obama's and Mitt Romney's Online Campaigns." In *The United States Presidential Election 2012: Perspectives from Election Studies, Political and Communication Sciences*, edited by C. Bieber and K. Kamps, 331–361. Berlin: Springer.

Sidiropoulos, P., V. Mezaris, I. Kompatsiaris, H. Meinedo, and I. Trancoso. 2009. "Multi-Modal Scene Segmentation Using Scene Transition Graphs." In *Proceeding MM '09 Proceedings of the 17th ACM international conference on Multimedia*, 665–668. New York. doi:10.1145/1631272.1631383.

Stanfill, M. 2012. "Finding Birds of a Feather: Multiple Memberships and Diversity without Divisiveness in Communication Research." *Communication Theory* 22 (1): 1–24. doi:10.1111/j.1468-2885.2011.01395.x.

Stommel, M. 2010. "Binarising SIFT-Descriptors to Reduce the Curse of Dimensionality in Histogram-Based Object Recognition." *International Journal of Signal Processing, Image Processing and Pattern Recognition* 3 (1): 25–36.

Stommel, M., and O. Herzog. 2009. "SIFT-Based Object Recognition With Fast Alphabet Creation and Reduced Curse of Dimensionality." In *International IEEE Conf. on Image and Vision Computing New Zealand (IVCNZ)*, edited by D. Bailey, 136–141. New York: IEEE Press.

Teichert, J. 2011. "Visuelles Erkennen von Objekten mit Ausprägungsvarianzen. " PhD diss., University of Bremen.

Thomas, M. 2009. "Localizing Pack Messages: A Framework for Corpus-Based Cross-Cultural Multimodal Analysis." PhD diss., Centre for Translation Studies, University of Leeds.

Thomas, M. 2014. "Evidence and Circularity in Multimodal Discourse Analysis." *Visual Communication* 13 (2): 163–189. doi:10.1177/1470357213516725.

Thompson, H. S., and D. McKelvie. 1997. "Hyperlink Semantics for Standoff Markup of Read-Only Documents." In *Proceedings of SGML Europe '97: The next decade – pushing the envelope*, 227–229.

Tseng, C. 2013a. "Analysing Characters' Interactions in Filmic Text: A Functional Semiotic Approach." *Social Semiotics* 23 (5): 587–605. doi:10.1080/10350330.2012.752158.

Tseng, C. 2013b. *Cohesion in Film: Tracking Film Elements*. Basingstoke: Palgrave Macmillan.

Tversky, B., J. Zacks, and B. Martin. 2008. "The Structure of Experience." In *Understanding Events: From Perception to Action*, edited by T. Shipley and J. Zacks, 436–464. Oxford: Oxford University Press.

Van Leeuwen, T. 1996. "Moving English: The Visual Language of Film." In *Redesigning English: New Texts, New Identities*, edited by S. Goodman and D. Graddol, 81–105. London: Routledge.

Vanhoutte, E., and R. Van Den Branden. 2010. "Text Encoding Initiative (TEI)." In *Encyclopedia of Library and Information Sciences (ELIS)*, edited by M. Bates and M. N. Maack, 5172–5181. New York: Taylor & Francis.

Wache, H., T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner. 2001. "Ontology-Based Integration of Information – a Survey of Existing Approaches." In *Proceedings of the IJCAI'01 Workshop on Ontologies and Information Sharing*, 108–117. Menlo Park, CA: American Association for Artificial Intelligence.

Wilson, R. C., E. R. Hancock, and B. Luo. 2005. "Pattern Vectors from Algebraic Graph Theory." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27: 1112–1124. doi:10.1109/TPAMI.2005.145.

Wright, W. 1975. *Sixguns and Society: A Structural Study of the Western*. Berkeley: University of California Press.

Zacks, J., and B. Tversky. 2001. "Event Structure in Perception and Conception." *Psychological Bulletin* 127 (1): 3–21. doi:10.1037/0033-2909.127.1.3.

Zacks, J. M. 2010. "How We Organize Our Experience into Events." *Psychological Science Agenda* 24 (4). Accessed 30 April 2016. http://www.apa.org/science/about/psa/2010/04/sci-brief.aspx