TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 12 (2020) 274–310

ISSN:1756-8765 online
DOI: 10.1111/tops.12476

This article is part of the topic "Visual Narrative Research: An Emerging Field in Cognitive Science," Neil Cohn and Joseph P. Magliano (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

# Computational Approaches to Comics Analysis

Jochen Laubrock,[a] Alexander Dunst[b]

[a]*Department of Psychology, University of Potsdam*
[b]*Department of English and American Studies, University of Paderborn*

Received 8 March 2019; received in revised form 17 August 2019; accepted 27 August 2019

**Abstract**

Comics are complex documents whose reception engages cognitive processes such as scene perception, language processing, and narrative understanding. Possibly because of their complexity, they have rarely been studied in cognitive science. Modeling the stimulus ideally requires a formal description, which can be provided by feature descriptors from computer vision and computational linguistics. With a focus on document analysis, here we review work on the computational modeling of comics. We argue that the development of modern feature descriptors based on deep learning techniques has made sufficient progress to allow the investigation of complex material such as comics for reception studies, including experimentation and computational modeling of cognitive processes.

*Keywords:* Comics; Computer vision; Deep learning; Document analysis; Scene perception; Eye tracking; Visual attention

## 1. Introduction

Comics are a fascinating topic of study for the cognitive scientist, as they make contact with research on both reading and visual attention and perception. Each panel of a comic is an abstracted visual scene, and thus the research tradition on visual perception, scene perception, and understanding is directly relevant (Greene & Oliva, 2009a, 2009b; Henderson & Hollingworth, 1999). At the same time, a verbal narrative is told in the

Correspondence should be sent to Jochen Laubrock, Department of Psychology, University of Potsdam, Karl-Liebknecht-Str. 24/25, 14476 Potsdam, Germany. E-mail: laubrock@uni-potsdam.de

captions, and dialogical elements as well as internal states of mind are represented in speech and thought balloons, respectively. The sequence of panels creates a narrative, which resonates in a mental model on the reader's side. This touches on research on reading (Rayner et al., 2012), working memory (Baddeley & Hitch, 1974; Oberauer & Kliegl, 2006), sequence processing (Reber, 1967), and language and narrative understanding (Gernsbacher, 1990; Jurafsky & Martin, 2009; Zacks & Tversky, 2001; Zwaan, 2016), as on the interaction of attention with these processes (Engbert et al., 2005, 2015; Laubrock et al., 2013; Schütt, Rothkegel, Trukenbrod, Engbert, & Wichmann, 2019; Tatler et al., 2017). Loschky et al. (2018) sketch an integrative framework that describes how these psychological "front-end" and "back-end" processes might work together during comics reading. Often, text and image need to be read together to understand the full story, with interesting implications for image-text interactions. Furthermore, the visual scenes in comics use a specific visual language with genre-specific syntactic and semantic features, vocabulary, morphology, and narrative structure (Cohn, 2013c).

Partly owing to this complexity, we are only beginning to understand the cognitive processing of comics (see Cohn, 2013c and the edited volume by Dunst et al., 2018 for recent attempts). An objective empirical description of complex visual stimuli in terms of linguistic and visual features is an important first step in studying their cognition. This article examines the current state of the art of computational comics analysis, which provides such a characterization and is therefore highly relevant for fields such as scene perception. The most promising description builds on features of deep neural networks, which are firmly rooted in cognitive science (Rumelhart & McClelland, 1986) but are now more commonly used for applied perception in engineering and computer science.

We begin by briefly introducing the surface features of comics and evaluate the data sets that are currently available for training computational models. The review of computational approaches that follows focuses on visual structure and text before turning to text–image combinations and the analysis of narrative structure. While focusing on comics analysis, these sections also illustrate modern techniques in digital image and natural language processing (NLP). The article ends with a brief section on the reading of comics, including eye tracking studies, some of which relate computational features to human behavior. The relationship of deep neural network feature descriptors, neuronal activity in the brain, and human behavior is a promising research topic with implications well beyond the study of comics.

## 2. The structure of comics

Comics are a complex type of document. They are usually published as printed magazines and books or included in other publications such as newspapers. The latter often contain short comic strips, whereas comic books can tell complex stories extending over many pages, each of which is composed of one or more individual images that are arranged in a layout and are frequently bounded by drawn visual borders (Fig. 1). These so-called panels are a fundamental structural and semantic unit in comics. They usually

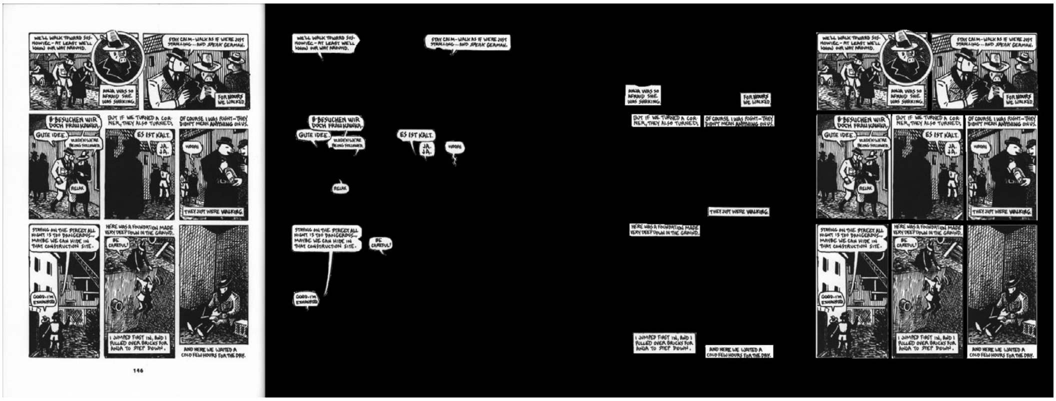
WE'LL WALK TOWARD SOSNOWIEC–AT LEAST WE'LL KNOW OUR WAY AROUND.
ANJA WAS SO AFRAID SHE WAS SHAKING.
STAY CALM–WALK AS IF WE'RE JUST STROLLING... AND SPEAK GERMAN.
FOR HOURS WE WALKED
B-BESUCHEN WIR DOCH FRAU KAWKA.
GUTE IDEE.
VLADEK-WE'RE BEING FOLLOWED.
RELAX
BUT IF WE TURNED A CORNER, THEY ALSO TURNED.
ES IST KALT.
JA. JA.
OF COURSE I WAS RIGHT–THEY DIDN'T MEAN ANYTHING ON US.
WOOF!
THEY JUST WERE WALKING.
STAYING ON THE STREET ALL NIGHT IS TOO DANGEROUS... MAYBE WE CAN HIDE IN THAT CONSTRUCTION SITE.
GOOD. I'M EXHAUSTED.
HERE WAS A FOUNDATION MADE VERY DEEP DOWN IN THE GROUND.
BE CAREFUL!
I JUMPED FIRST IN, AND I PULLED OVER BRICKS FOR ANJA TO STEP DOWN.
AND HERE WE WAITED A COLD FEW HOURS FOR THE DAY.
146

Fig. 1. Comic page (Spiegelman, 1997, p. 146) with ground truth annotations for speech balloons, captions, and panels (from left to right). Excerpt from COMPLETE MAUS by Art Spiegelman. Copyright © 1973, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1989, 1990, 1991, 1997 by Art Spiegelman, used by permission of The Wylie Agency LLC.

contain line drawings that depict a visual scene, including characters, objects, and background, as well as textual objects such as narrative captions and speech balloons, which represent narrative text and a given character's speech or thoughts, respectively. To this may be added other elements such as onomatopoeia that visualize sound, diegetic text, and motion lines. Finally, there are elements symbolizing meaning such as gears spinning above a character's head to convey that they are thinking. Cohn (2013b) has argued that these are lexical items of a visual languages, which can be used in a combinatorial manner. Such "visual morphemes" include "pictorial runes" (Forceville, 2005) and culture-specific elements, such as the "manpu" in Japanese manga, small symbols that express feelings and sensations of characters.

The composition and spatial arrangement of panels on a page into a page layout has many artistic degrees of freedom. Complexity of layout varies from a basic grid arrangement, via staggering—in which panels are slightly displaced horizontally or vertically—to more complex layouts that employ background images, insets, bleeding, and panels defined by illusory contours (Cohn, 2013a). Like other aspects of comics storytelling, the complexity of page layouts has increased over historical time (Pederson & Cohn, 2016). Nonetheless, page layout needs to be navigable so that panels can be read in the intended order, partly governed by cultural conventions.

In fact, the sequential structure of comics is one of their defining features, which is why comics have been called sequential art (Eisner, 1985; McCloud, 1993). The empty space between panels is called the gutter, and the transition from panel to panel requires an act of inference from the reader. Therefore, panel transitions can be categorized by the type and amount of inference required. In his well-known comic about comic art, McCloud (1993) defines such a taxonomy and presents one of the earliest corpus analyses of panel transitions. The amount of inference required is strongly tied to the semantic content of panels, making computational analysis of panel transition a difficult problem.

In addition, the comics reader has to create a mental model of the story beyond simple panel-to-panel transitions. Cohn (2010) points out that such construction of narrative sequences in comics is guided by hierarchical structures (much like in language, music, or vision) and provides an in-depth analysis of these higher-order structures (Cohn, 2013b).

Cohn and Magliano (this volume; see also Cohn, 2013b) propose a framework of the structures involved in visual narratives: A modality of expression (graphics) is used to express conceptual information (meaning) by employing combinatorial systems (grammar). These structures work at both the unit and sequence levels. For example, each single panel employs a visual vocabulary, and panel layout organizes the reading sequence. At the conceptual level, visual narratives convey meanings, both within panels and across sequences, requiring the construction of situation and event models on the reader’s side. Conventionalized visual morphological features similar to the morphology of a language are used to construct units, and narrative structure is organized by a narrative grammar that resembles syntax in a more standard language. Given the complexities of comics and their unique visual language, their computational analysis is an interesting but challenging problem.

## 3. Computational approaches

While comics have been studied in the humanities for several decades now, computational approaches to comics analysis are still in their infancy. This means that several of the categories proposed in Cohn and Magliano’s framework have barely been touched by computational approaches. Existing research mostly deals with presentation aspects at the unit level, that is, graphic and morphological structure. Higher-level conceptual structure is still a relatively neglected topic in computational comic analysis, as are most sequential aspects, especially narrative and event structure.[1] So the term *infancy* from our initial sentence may be taken quite literally to mean that the computational analysis of comics at present—like an infant—does not address referential continuity. This is partly due to the complexity of the medium: A computational analysis of comics requires computer vision to understand page layout, visual scenes inside and beyond individual panels, and characters, objects, actions, and emotional expressions. In turn, natural language processing needs to process story and dialogue, multimodal representations for text-image relations, and create and maintain a high-level narrative model to capture events, narrative arcs, inter-character relations, and so on.

Computational approaches to comics can be broadly partitioned into content analysis, content generation, and user interaction (Augereau et al., 2017). In this review, we mainly focus on content, specifically, document and narrative analysis. We also touch on user interaction, with a focus on which elements of comics are attended, as measured by gaze locations in eye-tracking studies and predicted by computational models of saliency. We will start by briefly describing the data sets that are currently available for evaluating computational models. We then summarize results from document image analysis, including visual stylometry, semantic segmentation, text recognition, and more generally, locating, segmenting, and identifying relevant objects and characters. Both classical feature engineering and modern deep learning approaches to image and text analysis will be discussed. Due to the limited availability of training data, deep learning has come relatively late to comics analysis, and these lower-level document analysis tasks have only very recently made sufficient progress to provide input for higher-level analyses. Thus, we will finish by covering the few attempts that have been made to computationally analyze narrative structure. Combining existing theoretical conceptions like Cohn's (2013c) visual narrative grammar with available data sets and computational analysis has the potential to provide exciting new tools for the comparative study of comics.

Given the fast pace of development in computational methods following advances in training deep neural networks for vision and language understanding, the present review necessarily only provides a snapshot. Readers who want to remain up-to-date on these developments are encouraged to follow the proceedings of the International Workshop on coMics ANalysis, Processing and Understanding (MANPU; Ogier et al., 2016) as well as conference proceedings devoted to more general topics in computer vision, computational linguistics, document analysis, and artificial intelligence (such as NIPS, CVPR, ECCV, ICCV, ACL, and ICDAR).

## 4. Data sets

Until recently, no corpora of digitized comics were available for research. Early studies (Manovich, 2012; Saito & Matsui, 2015) therefore relied on scraping very large image boards of debatable legal status such as Manga "scanlation" sites.

In the past few years, some research corpora have been established.[2] These can be divided into data sets in which the image scans themselves are accessible (Manga109, BCD/COMICS, eBDthèque) because the copyright has expired or has been granted for research purposes, and others (GNC, VLRC) for which metadata and image descriptions at various levels of resolution may be available, but the scans themselves cannot be distributed because of copyright restrictions. Obviously, the public data sets are more useful, but some of the publicly accessible information in private data sets is unique. Table 1 gives an overview of the available data sets, which are described in more detail below.

### *4.1. eBDthèque*

The L3I group at La Rochelle provided the first publicly available data set, eBDthèque (Guérin et al., 2013). The eBDthèque contains 100 pages from 25 different French, American, and Japanese comics published between 1905 and 2012. The data set also includes manual annotations of the panel locations, text lines, balloons, reading order, balloon shape (pixel-based in v2), and text transcription, as well as metadata on album title, artist, release date, and so on. Heterogeneity in style was a guideline in curating this corpus. Balloons, for instance, may "be closed or half-closed, oval, rectangular, peaky, wavy, with or without tail and with a white or yellow background" (Guérin et al., 2013, p. 2).

Table 1
Comics data sets

| Data set | Topic | Source Decade | No. Books | No. Pages | No. Pages (annotated) | Images Public | Annotations | Text Transcription | Reference |
|---|---|---|---|---|---|---|---|---|---|
| eBDthèque | French, American, Japanese comics | 1900–2010 | 25 | 100 | 100 | Yes | Panels, text lines, balloons, balloon shape (polygons) | Yes | Guérin et al. (2013) |
| Manga109 | Japanese manga | 1970–2010 | 109 | 21,142 | 21,142 | Yes | Panels, text boxes, characters, faces | Yes | Matsui et al. (2017) |
| COMICS | American golden age comics | 1930–1950 | 3,948 | 198,657 | 198,657 | Yes | Panels*, text boxes* | Yes* | Iyyer et al. (2017) |
| DCM_772 | American golden age comics | 1930–1950 | 27 | 772 | 772 | Yes | Panels, characters, faces | Partly* | Nguyen et al. (2018) |
| VLRC | Comics from various places and genres | 1940–2010 | 292 | 6,200 | 6,200 | No | Panels, semantic relations between panels, page layout, multimodality | No | Cohn et al. (2017) |
| GNC | English-language graphic novels | 1970–2010 | 270 | 55,000 | ca. 3,000 | No | Panels, balloons, captions, onomatopoeia, diegetic text, characters (all polygons) | Yes | Dunst et al. (2017) |

*Note*. Object annotations are bounding boxes if not otherwise specified. An asterisk (*) indicates machine annotations (or human-in-the-loop).

Given its relatively small size, the eBDthèque is not really suitable as training material for machine learning approaches but represents a very challenging evaluation set due to its heterogeneity.

### 4.2. Manga109

The Manga109 data set (Matsui et al., 2017) was compiled by the Aizawa Yamasaki Laboratory at the University of Tokyo. Manga109 is composed of 109 manga volumes from a wide range of genres, with a total of 21,142 pages drawn by professional manga artists in Japan between the 1970s and 2010s. Permission from 94 professional creators was obtained to make the data set publicly available for academic use. Ogawa et al. (2018) added verified ground truth for each page, consisting of an impressive 527,685 human annotations of bounding boxes in four categories: panel/frame, text object (e.g., caption, speech balloon, onomatopeia), face, and body. Furthermore, character names (identities) and text content have been manually annotated. With about 100 times as many annotations per category as eBDthèque, Manga109 excels both in size and annotation quality, making it a good candidate for machine learning training material. However, the data set is limited to Manga, which has its own unique stylistic elements and visual vocabulary. As a consequence, results obtained by training on Manga109 will not readily transfer to comics from US or European traditions.

### 4.3. DCM/COMICS

The Digital Comic Museum (digitalcomicmuseum.com, DCM) is a digital library of golden-age comic books in the public domain, published before 1959.[3] Iyyer et al. (2017) created the COMICS corpus by downloading the 4,000 highest-rated comic books from DCM. They semi-automatically segmented each page into panels, extracted textbox locations from the panels, and transcribed the text by using a commercial OCR service on the resulting 2.5 million images of textboxes. Faster R-CNN (Ren et al., 2015) networks were trained for bounding box segmentation of panels and textboxes given a limited number of manually annotated examples (500 pages, 1,500 textboxes). The annotated data set is available for download at https://github.com/miyyer/comics. Whereas the size of the data set is impressive, the fact that most of the segmentation has been done by machines limits the data quality.

Another subset of the DCM was fully annotated by humans. Nguyen, Rigaud, and Burie (2018) created DCM_772, a subset of 772 pages from the Digital Comic Museum stratified by publisher. Ground truth is available in the form of object bounding boxes around characters, which are further differentiated into four classes: human-like, object-like, animal-like, and extra (supporting role characters), making this data set potentially suitable for differentiated character detection and recognition.

In addition to these public corpora, we would like to mention two private corpora which have been used in previous work. In contrast to the above-mentioned data sets, they were curated by experts from literary studies or cognitive science.

### *4.4. VLRC*

Neil Cohn has compiled the Visual Language Research Corpus (VLRC), which according to the website http://www.visuallanguagelab.com/vlrc/ (retrieved March 7, 2019) consists of

> 35,000 coded panels from roughly 300 comics from several places (America, Belgium, France, Germany, Hong Kong, Japan, Korea, The Netherlands, Sweden), different time periods (1940–present), and various genres. It includes coding of panel framing, semantic relations between panels, external compositional structure (page layout), multimodality, and a variety of other structures of visual languages.

At least 25 pages or 150 panels per comic were annotated by human coders, with roughly 60% of the comics annotated by two coders each. Cultural diversity across visual languages is an advantage of the VLRC, which has been used in a number of studies on page layout, multimodality, semantic structure, and narrative structure (e.g., Cohn, 2019; Cohn, Pederson, & Taylor, 2017; Pederson & Cohn, 2016).

### *4.5. GNC*

The Graphic Narrative Corpus (GNC; Dunst et al., 2017) has been curated with the specific goal to create a data set useful for research in the humanities, aiming for representativeness and attempting to balance popularity and prestige. The GNC focuses on English-language graphic novels and currently contains about 270 titles with approximately 55,000 pages. A majority of the titles was published in the United States and Canada, but the corpus also contains graphic novels from other regions. All works have a length of at least 64 pages, tell a continuous story, and target an adult readership. The GNC contains both fictional and non-fictional texts from the 1970s to the present. Human annotations of selected pages of each volume include polygons around panels, main characters, speech balloons, captions, onomatopoeia, and diegetic text, as well as transcriptions of all text objects. Metadata include bibliographical information, translations, adaptations, prizes awarded to individual books, and a cast list of all characters. The original page images cannot be distributed, but page descriptions in the form of CNN activations are available. A unique feature of the GNC are eye movement recordings of at least 10 readers of the initial pages of each volume. The research group responsible for the GNC has also developed the free M3 annotation software, https://groups.uni-paderborn.de/graphic-literature/wp/?page_xml:id=3592. Empirical studies using the GNC include work on illustrator classification, semantic segmentation, and eye movement modeling (Dubray & Laubrock, 2019; Laubrock & Dubray, 2018, 2019; Laubrock, Hohenstein, & Kümmerer, 2018).

## 5. Analysis of visual structure

Cohn (2013b) provides a detailed overview and description of what the individual constituents of visual structure are. Walsh created Comic Book Markup Language

(CBML), a TEI-based XML vocabulary (with DTD and schema representations) for the XML encoding of comic books and graphic novels. Whereas CBML is focused on textual aspects of comics, several alternatives or extensions have been proposed that allow polygon-based annotations of visual objects (Dunst et al., 2017; Guérin et al., 2013).

In what follows, we will describe attempts to analyze, detect, and segment several of the elements described above with the help of computational methods. The reader may bear in mind the conclusion reached by a recent review, that "segmenting the panels or reading the text of any comics is still challenging because of the complexity of some layouts and the diversity of the content," "the segmentation of faces and body of the characters is still an open problem," and "some parts of comics has not been addressed at all, such as the detection of onomatopoeias" (Augereau et al., 2017).

### *5.1. Computer vision: Feature engineering and deep learning*

Before we address the analysis of comics proper, we will briefly summarize the state of the art in computer vision approaches to document analysis. Since Krizhevsky et al. (2012), we have witnessed a deep-learning revolution in the computational analysis of visual stimuli (LeCun et al., 2015). Traditional computer vision techniques require careful design to arrive at engineered feature extractors like edge-oriented histograms (EOH; Levi & Weiss, 2004), histograms of oriented gradients (HOG; Dalal & Triggs, 2005), or SIFT (Lowe, 2004) that transform the raw input (e.g., pixels of an image) into a higher-level representation useful for detection, segmentation, or classification. However, feature engineering is difficult and expensive, as it requires domain knowledge and engineering expertise. Furthermore, even the aforementioned highly integrated features are still relatively low-level in terms of a visual object hierarchy.

Deep-learning methods, in contrast, learn useful representations from the data by adjusting weights of feature hierarchies. The hierarchical processing architecture is inspired by biological visual systems. Features at the lower levels are relatively close to the stimulus, whereas they become increasingly abstract and object-like at higher layers. Importantly, these feature layers are not designed by a researcher but learned from the data using error backpropagation/gradient descent.

Fully connected neural networks (Fig. 2, top) have a large number of parameters and are therefore hard to train and prone to overfitting when they are more than a few layers deep, because each unit of one layer is connected to every unit of the next. The so-called convolutional neural networks (CNNs; LeCun et al., 1989; Fig. 2, bottom) are much more sparse and hence easier to train and more regularized. CNNs achieve this sparsity by using local connections, shared weights, pooling, and the use of many layers. Convolutional layers in a CNN are simply banks of small filters for local image patches, in which the filter coefficients are learned. A given filter is applied across the whole image, independent of location, dramatically reducing the number of parameters. At the lowest level, these filters describe simple regularities in the image.

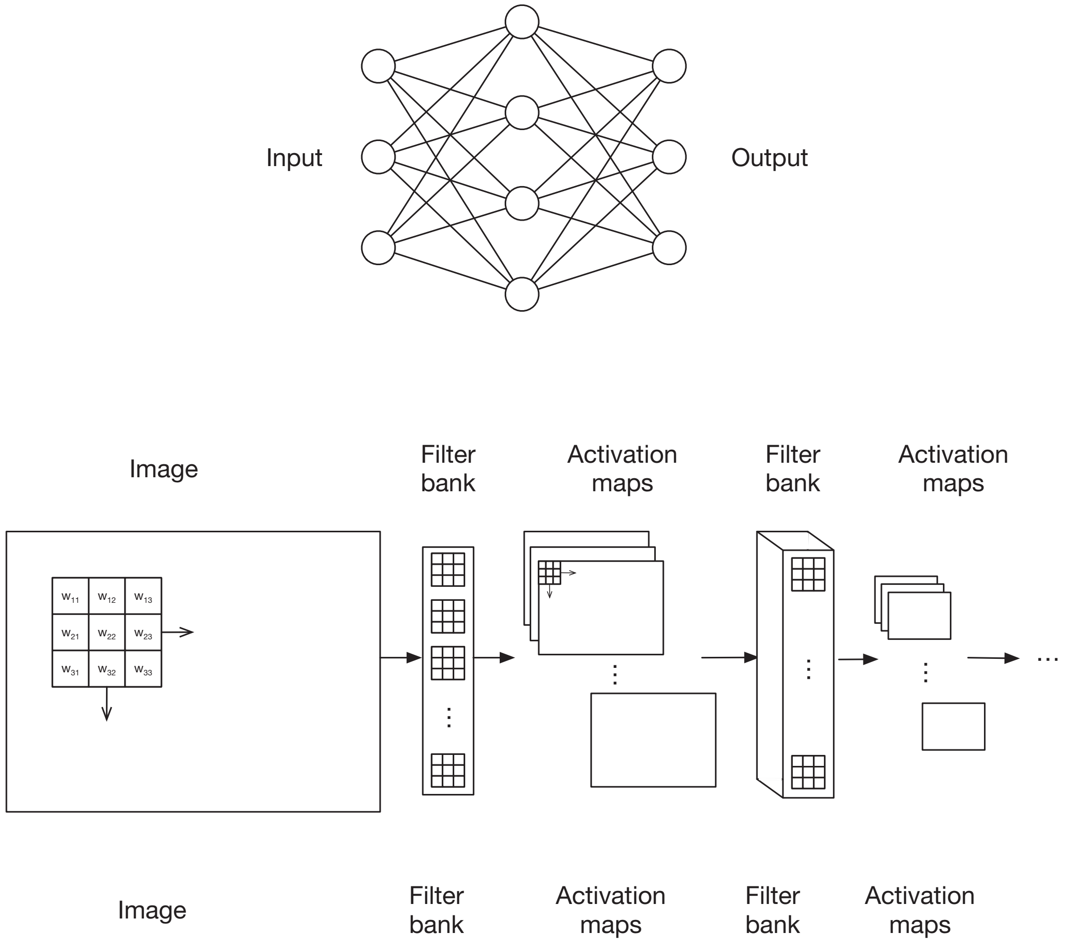

Fig. 2. Fully connected (top) and convolutional (bottom) neural network architectures.

At higher levels, they capture conjunctions of lower-level features. Interspersed max-pooling layers increase robustness by providing translation invariance and also reduce the spatial size of the representations, which reduces computational cost and makes higher-level representations increasingly coarse-grained. The motivation for the design of CNNs comes from the observation that natural signals can often be thought of as compositional hierarchies, which is one reason that biological signal processing systems such as human perception have evolved hierarchical processing stages.

> Deep neural networks exploit the property that many natural signals are compositional hierarchies, in which higher-level features are obtained by composing lower-level ones. In images, local combinations of edges form motifs, motifs assemble into parts, and parts form objects. Similar hierarchies exist in speech and text from sounds to phones, phonemes, syllables, words and sentences. [. . .]

> The convolutional and pooling layers in ConvNets are directly inspired by the classic notions of simple cells and complex cells in visual neuroscience, and the overall architecture is reminiscent of the LGN–V1–V2–V4–IT hierarchy in the visual cortex ventral pathway.
>
> —(LeCun et al., 2015, p. 439)

CNNs (or ConvNets) are now the dominant approach for almost all recognition and detection tasks and surpass human performance on some tasks. Representations at the lower layers of CNNs are quite generic, and models pre-trained on large-scale image or object classification tasks such as ImageNet (Deng et al., 2009), PASCAL VOC (Everingham et al., 2015), and MS COCO (Lin et al., 2014) are readily available.

### *5.2. Object recognition, semantic segmentation, and instance segmentation*

For the visual analysis of comics, the related computer vision tasks of object recognition, semantic segmentation, and instance segmentation are particularly important. Object recognition requires finding all objects from a set of object categories in an image, and localizing them with their bounding boxes. Semantic segmentation classifies each pixel of the image as belonging to one of several categories, which typically results in a much finer delineation of object shapes. For example, the output of semantic segmentation might be a set of masks for speech balloons, captions, and panels. In instance segmentation, additionally, different instances of the same class are separately labeled. In the example, each individual speech balloon would be assigned an individual label.

Two classes of models currently dominate the object recognition literature: Models from the R-CNN family (Girshick, 2015; Girshick, Donahue, Darrell, & Malik, 2014; Ren et al., 2015) use a two-stage process. First, a set of object candidate regions are proposed, and CNN features are computed for these region proposals. Second, the feature vectors are used as regressors to predict object bounding boxes and class probabilities. In contrast, the so-called one-stage object detection models such as YOLO (Redmon et al., 2015, 2017), SSD (Liu et al., 2016), or FPN (Lin et al., 2017) skip the region proposal stage and run detection directly over many possible locations. Thus, a single neural network is applied to the full image. Whereas single-shot models are much faster, two-stage models tend to achieve somewhat higher accuracy.

At some stage, fully connected layers are needed for all of the above object recognition models. Interestingly, the seemingly more difficult task of semantic segmentation has been dominated by fully convolutional approaches for a while, that is, networks that use only convolutional layers (e.g., Long, Shelhamer, Darrell, & 2015). These typically consists of a "bottom-up" encoding and a "top-down" decoding branch. A standard CNN such as the convolutional part of a VGG model (Simonyan & Zisserman, 2014) can be used to encode the image by a hierarchy of learned filters coding for increasingly semantically rich features with decreasing spatial resolution. The decoding branch performs intelligent up-sampling by transposed convolutions, again with learned filters, and often integrates a copy of the encoder representation at the corresponding level using skip connections, as in the U-Net (Ronneberger et al., 2015) or FPN (Lin et al., 2017).

Instance segmentation as a task combines aspects of object recognition and semantic segmentation: The location and identity of objects are to be predicted, as well as an exact pixel-wise segmentation mask. One of the current state-of-the-art models, Mask R-CNN (He et al., 2017), thus combines ideas from the R-CNN approach with a fully convolutional FPN-based backbone.

A potential problem with applying deep learning to comics analysis is that most of the readily available pretrained models were trained on images of natural scenes rather than illustrations. Training models with many parameters from scratch requires large amounts of data—for example, the MS COCO data set (Lin et al., 2014) has 2,500,000 labeled instances in 328,000 images. As we have seen, annotated data sets of this size are not available for comics.

However, good performance in classification of illustrator style is possible using models pretrained on photographs (Laubrock & Dubray, 2019), suggesting that their lower- and mid-level features are generic enough to also describe comics. In general, learned CNN filters at the lower layers of the visual hierarchy are quite generic, so that they can be reused for many different visual tasks. By fine-tuning, that is, only retraining a few of the upper layers, good performance and transfer to new tasks and materials can often be achieved with limited data sets containing relatively few training samples. Using this procedure, deep neural networks have successfully been applied to comics analysis.

Indeed, the state of the art in visual document analysis of comics is dominated by approaches based on deep CNNs. Given their data demand and the fact that most of the data sets described above have only been available very recently, these results may come as somewhat of a surprise, and illustrate the potential of these approaches. Table 2 summarizes the categories which we will discuss in detail below.

Table 2
State of the art in visual analysis of comics

| Task | References | Remarks | *F*-score |
|---|---|---|---|
| Panel segmentation | Rigaud et al. (2015), Ogawa et al. (2018), and Laubrock and Dubray (in press) | Mostly just bounding boxes rather than exact panel shape | >0.9 |
| Page layout | Cohn (2014), Bateman et al. (2017), and Tanaka et al. (2007) | Mostly conceptual, but should work fine with current panel segmentation techniques | N/A |
| Caption and balloon segmentation | Ogawa et al. (2018), Nguyen et al. (2019), and Dubray and Laubrock (2019) | Distinction between captions and balloons only in Laubrock/Dubray | >0.9 |
| Character and object detection | Matsui et al. (2017), Qin et al. (2017), and Nguyen et al. (2019) | Mostly character detection | >0.6 |
| Stylometry | Laubrock and Dubray (2019) | Artist classification easier than genre classification | >0.9 |

### 5.3. Panels

Most comics have panels with clearly outlined borders and gutters, which have been compared to the punctuation and spacing in written and printed texts (Walsh, 2012). Panels of this type are relatively easy to detect with the help of classical approaches like connected-component labeling (Rosenfeld & Pfaltz, 1966). For example, Rigaud, Tsopze, Burie, and Ogier (2013) proposed a method that computes connected components of a binarized page and classifies the extracted component as noise, text, or panel according to their height with the help of a *k*-means cluster analysis with $k = 3$. Post-processing uses topological relations for removing false positives (e.g., a panel should not be fully contained in another panel). Rigaud, Guérin, Karatzas, Burie, and Ogier (2015) refined the approach by adding a few preprocessing steps, and adding the final step of computation of the convex hull of the panel contours in order to recover from discontinuous contour detection. This heuristic method works fairly well for standard cases (and indeed it still represents the state of the art in panel segmentation as evaluated with the eBDthèque) but fails in others. For example, it will erroneously filter inset panels and encounter problems with panels that are connected by objects in the foreground. Similarly, a complex algorithm that uses a set of engineered features developed by Stommel et al. (2012) can only detect rectangular panels.

However, panels without clear borders and gutters are common, as are panels with irregular shapes. Pang et al. (2014) list several problems for traditional approaches to panel segmentation, such as non-rectangular panel shapes, parts of foreground objects lying outside the panel or connecting two panels, and panel borders partly or completely defined by illusory contours. They developed a method that can handle these cases, using the three steps of panel block generation, panel block splitting, and panel shape extraction. Illusory contours are converted to real contours by some clever heuristic tricks. Once again, the method starts with connected component labeling and uses clever heuristics and prior knowledge of comics layout. At the last step, panel shape is recovered by detecting the four corners of each panel. The authors achieved impressive results on manga with complex layouts, showing that the method is relatively robust against stylistic variations. An earlier method proposed by Tanaka et al. (2007), which is based on a detection of straight lines by what appears to be a variant of the Hough transform, followed by an evaluation of T-junctions, has similar qualities.

Although the approaches by Tanaka et al. (2007) and Pang et al. (2014) deliver good results, they require pages with (more or less) uniformly white backgrounds and clean gutters—a fact they have in common with most other classical feature-engineering approaches. For this reason, Iyyer et al. (2017) took an alternative route and used deep learning to analyze page scans from the DCM. They annotated 500 randomly selected pages with rectangular bounding boxes for panels and then trained a Faster R-CNN to automatically detect panels. This seems like a promising approach but is hard to evaluate because Iyyer et al. (2017) neither included measures of their panel segmentation performance nor an evaluation comparing their approach to earlier methods.

Nguyen et al. (2018) used an off-the-shelf YOLOv2 region proposal network (Redmon & Farhadi, 2017) with some adjusted priors to detect bounding boxes around panels, characters, and faces. For panels, they achieved comparable or slightly better performance than feature engineering techniques, especially when the testing material was from the same set of comics books as the training material. However, they conclude that overall, traditional methods still work slightly better. Nguyen et al. see one advantage of deep-learning based models in their ability to correctly detect difficult panels with a complex background where existing methods fail.

Both Nguyen, Rigaud, and Burie, (2019) and Ogawa et al. (2018) present CNN-based models that can detect multiple objects at the same time, including panels. The Nguyen et al. (2019) model is described in more detail in the section on speech balloons. The model by Ogawa et al. (2018) can predict panels, character bodies, faces, and text regions. It is based on a SSD300 object detection model (Liu et al., 2016), which the authors modified to address the comics-specific problem of multiple overlapping objects of similar location and size (e.g., a zoomed-in character that nearly fills the panel). The model makes very good panel detections for the test set of the Manga109 data set, from which its training set was taken. Predictions for the eBDthèque, which is very heterogenous in panel style, are on a par with Ngyuen et al. (2019), but decidedly weaker than Rigaud, Guérin, et al. (2015).

Thus, a method based on classical engineered features combined with heuristics capturing expert knowledge still works best on the eBDthèque, possibly suggesting that the data sets used for training CNN-based models are just not diverse enough to learn generalizations applicable across the wide variability of artistic styles of the eBDthèque. It remains to be seen how CNN-based approaches using more diverse training sets can compete.

### *5.4. Page layout*

Comics artists spend considerable time thinking about how to compose panels across a page, and page layout is characteristic of an illustrator's style (Laubrock & Dubray, 2019). Comics layouts can range from typical and simple n-by-m grids of equal-sized panels to rather complex arrangements with different-sized panels, oblique or curved borders, insets and bleeding. Layout analysis is further complicated by the fact that panel boundaries are often broken up by text elements such as speech balloons or onomatopoeia as well as graphical elements, or are even borderless. Furthermore, the background sometimes serves as another panel, adding a depth layer, and gaps can occur in the regular grid layout. Panels are also sometimes grouped together to form larger units, often using perceptual principles of grouping. As Walsh (2012) puts it, "Panels of every size and shape—circles, triangles, irregular polygons, and fluid organic shapes—are also common, as are borderless panels with ambiguous boundaries." Panel layout co-determines reading order, which normally follows cultural conventions and might differ between languages: Western comics are read from left-to-right, top-to-bottom, whereas for Japanese comics the reading order is usually right-to-left, top-to-bottom.

However, sometimes even the sequential ordering of panels is unclear, although this is usually limited to short sequences or experimental comics (e.g., Ware, 2012). On the one hand, the standard grid layout leading to a Z-path sequence in Western culture is very common, allowing Iyyer et al. (2017) to use used a simple Morton order (Morton, 1966) of the midpoints of their automatically segmented panels, noting that incorrect orderings occurred only for rare and complicated panel layouts. On the other hand, capturing all eventualities might be impossible, and even describing cases that are not extremely uncommon requires a complex taxonomy.

Tanaka et al. (2007) describe a tree-based method that might not cover all cases, but can handle the somewhat complex layouts typically employed in Manga. Another tree-based description of layout was proposed by Cohn (2013b, 2014), which also accounts for variations like inset panels. This system has been used for comics annotation in studies on cross-cultural differences and the historical development of layouts (Cohn et al., 2017; Pederson & Cohn, 2016).

Bateman et al. (2017) have developed a comprehensive formal classification scheme, inspired by the potential narrative relevance of panel organization. The scheme can be used for annotation and is ready for computational implementation (it has been integrated into the M3 editor mentioned above).

In summary, a formal description of panel layout across a page is well-developed and has been used in annotation. Assuming a flawless panel detection that returns a description in terms of geometric panel properties, rule-based classification of layout of the vast majority of comics pages should be possible with the aid of these systems. In fact, some systems even implement automatic page layout, given panel properties (Cao, Chan, & Lau, 2012; see the description at the end of this review).

### *5.5. Captions and balloons*

Narrative text in comics is often placed in captions, which tend to be rectangular and located near the upper or lower panel boundary. Balloons, in contrast, are mostly used to convey direct speech. Speech balloons often include pointed tails originating from the speaker. Similarly, thought balloons express a character's thoughts and often include a path of little circles originating from the thinker. A third type of balloon is "audio" balloons emanating from radios, television sets, loudspeakers, and the like, usually characterized by jagged borders. Forceville et al. (2010) provide a more comprehensive description of balloons. Cohn (2013b) points out that the visual surface features of balloons are not always aligned with meanings such as speech or thoughts, again attesting to artistic freedom.

Reading order of balloons within a panel typically follows the same rules as reading order of panels on a page. Detecting and segmenting captions and balloons is not only interesting for describing the comic but also an important step in a pipeline for OCR, which works much better with isolated text than with text on complex pages, especially since the visual textures used in comics are often in the same spatial frequency band as letters. Interestingly, most approaches to detection and segmentation of captions and speech balloons have not tried to differentiate between the two, despite their different narrative functions. Possibly this is due to limitations in the annotations, which often lump them together as text objects. Corpora like the GNC that honor the different classes or additional post-processing should therefore be used when narrative function is a topic of analysis.

In their related approaches to speech balloon detection, Arai and Tolle (2011) and Ho et al. (2012) use connected component labeling, some morphological operations such as erosion and dilation, and heuristics that include size and height relative to panel dimension and white pixel ratio. These approaches work better than chance, but are not ready for production use (see Nguyen et al., 2019, Table 1). Rigaud and colleagues have produced several other models for speech balloon detection and segmentation based on feature engineering. Their connected-components based approach (Rigaud, Burie, & Ogier, 2017) has a higher detection rate because it uses an adaptive threshold combined with text analysis. An advantage of this method is that it gives pixel-based contours, making it possible to identify the direction of the tail, which can be used to associate speech balloons and comic character (Rigaud, Guérin, et al., 2015; Rigaud, Thanh, et al., 2015). However, all of these methods only work with closed balloons.

Rigaud, Burie, Ogier, Karatzas, and van de Weijer (2013), realizing the great heterogeneity of balloons, which are often defined by irregular shapes or even illusory contours, developed a decidedly different approach based on an active contour model, which they further tailored to balloon detection based on domain knowledge, such as strong edges, smooth contours, and the relative location of text. This method can delineate the full shape of a speech balloon and handle illusory or subjective contours. However, it is relatively slow, relies on text spotting as a preprocessing step, and works less well than Rigaud et al. (2017) for detection. A combination of the two approaches appears promising for handling detection and segmentation.

Two recently proposed models based on CNN features (Nguyen et al., 2019; Dubray & Laubrock, 2019) achieve significantly better pixel-based speech balloon segmentation than those based on feature engineering. Additionally, Nguyen et al. (2018) obtained good results for speech balloon detection on the eBDthèque by using an off-the-shelf DeepLab semantic segmentation network (Chen et al., 2018).

Nguyen et al. (2019) describe a multi-task learning model for comic book image analysis derived from Mask R-CNN (He et al., 2017) that can handle detection and segmentation of several object categories, including speech balloons, panels, and characters, and also predicts character-balloon associations. Nguyen et al. (2019) added a binary classification branch to predict for each pairwise combination of detected balloons and characters whether they are associated or not. For balloon segmentation, they found that a plain Mask R-CNN as well as their adapted version called Comic MTL outperformed all traditional methods by a large margin on the eBDthèque. Comic MTL as a multi-task learning model has the additional advantage of simultaneously performing other tasks such as panel detection and character-balloon association. The SSD-based multi-task model by Ogawa et al. (2018) described above similarly achieves very good performance on the Manga109 corpus, but it only detects bounding boxes around text and cannot do pixel-wise segmentation.

Dubray and Laubrock (2019) use a different architecture based on a fully convolutional U-Net (Ronneberger et al., 2015) end-to-end encoder-decoder network. U-Net was developed for pixel-level semantic segmentation of medical images but transfers well to other semantic segmentation tasks. Laubrock and Dubray (in press) have extended the model to multi-class prediction of balloons, captions, and panels, achieving state-of-the-art results in each of these tasks (Fig. 3).

For encoding the visual scene, Dubray and Laubrock (2019) used a pretrained VGG-16 model (Simonyan & Zisserman, 2014), a conceptually simple encoder which they had previously found to have decent transfer to comics (Laubrock & Dubray, 2018; Laubrock et al., 2018). They added a decoding branch that performs intelligent upsampling combined with skip connections (Fig. 4). This architecture serves to fully recover the fine-grained spatial information lost in the pooling operations of the encoding branch. Dubray and Laubrock (2019) trained their model on the annotated GNC corpus and achieved very good segmentation performance on the test set, observing that the model could handle wiggly tails, curved corners, and illusory contours, and transferred well to the eBDthèque data set on which it had not been trained. Unlike most others, the model can distinguish between panels and captions due to its training on GNC annotations. Interestingly, Dubray and Laubrock (2019) report that the model initially did not work well on Manga109, suggesting that either characteristic aspects of the visual language, the difference in text orientation, or the writing system was used by the model in locating speech balloons. We have since then annotated a small sample of manga pages and retrained the model, and these difficulties are now overcome.

## 5.6. Characters and objects

Due to the relatively stereotypical appearance of panels, captions, and balloons, models for their detection and segmentation are nearly ready for production use in document

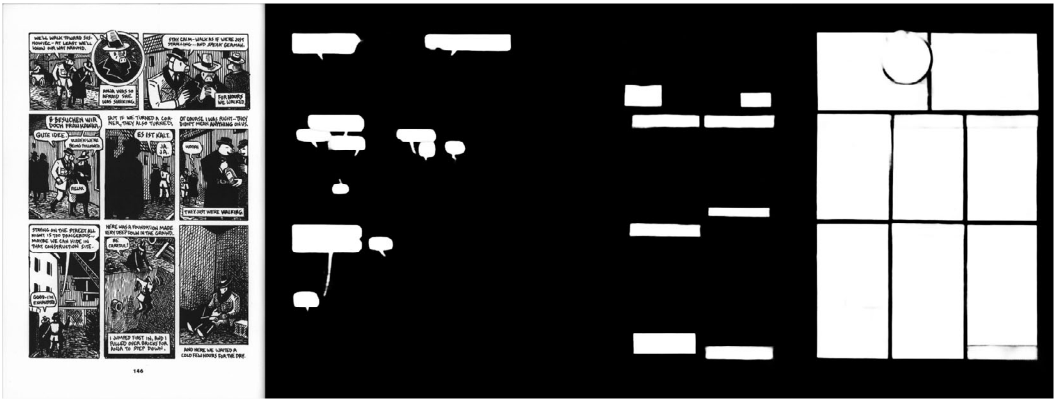
WE'LL WALK TOWARD SOSNOWIEC—AT LEAST WE'LL KNOW OUR WAY AROUND.
ANJA WAS SO AFRAID SHE WAS SHAKING.
STAY CALM—WALK AS IF WE'RE JUST STROLLING—AND SPEAK GERMAN.
FOR HOURS WE WALKED.
B-BESUCHEN WIR DOCH FRAU KAWKA.
GUTE IDEE.
VLADEK-WE'RE BEING FOLLOWED.
RELAX.
BUT IF WE TURNED A CORNER, THEY ALSO TURNED.
ES IST KALT.
JA. JA.
OF COURSE I WAS RIGHT—THEY DIDN'T MEAN ANYTHING ON US.
WOOH
THEY JUST WERE WALKING.
STAYING ON THE STREET ALL NIGHT IS TOO DANGEROUS... MAYBE WE CAN HIDE IN THAT CONSTRUCTION SITE.
GOOD—I'M EXHAUSTED.
HERE WAS A FOUNDATION MADE VERY DEEP DOWN IN THE GROUND.
BE CAREFUL!
I JUMPED FIRST IN, AND I PULLED OVER BRICKS FOR ANJA TO STEP DOWN.
AND HERE WE WAITED A COLD FEW HOURS FOR THE DAY.
146

Fig. 3. U-net based segmentation of a comics page (Spiegelman, 1997, p. 146) into speech balloons, captions, and panels (Laubrock & Dubray, in press). Excerpt from COMPLETE MAUS by Art Spiegelman. Copyright © 1973, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1989, 1990, 1991, 1997 by Art Spiegelman, used by permission of The Wylie Agency LLC.

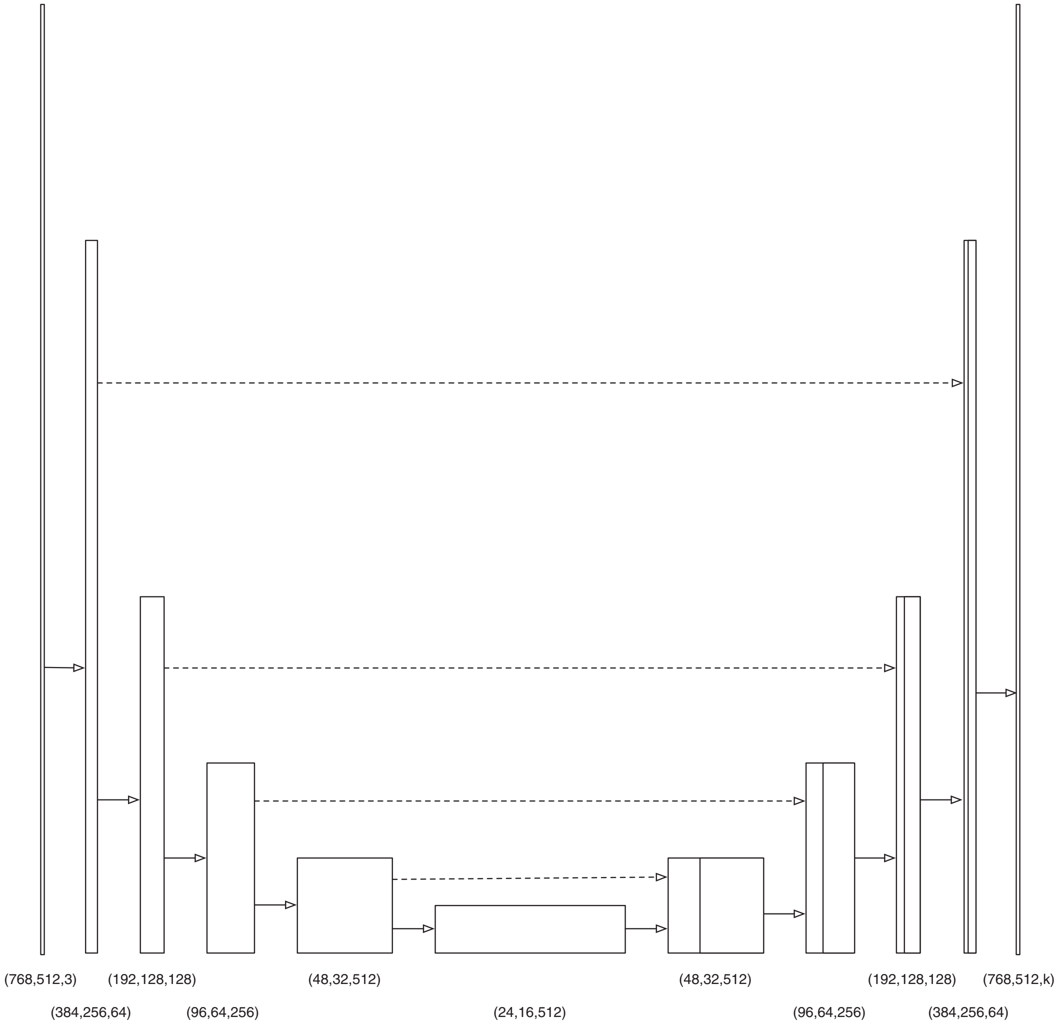

Fig. 4. Schematic representation of the Dubray/Laubrock U-net model. Information flow is from the input image on the left to the set of k output masks on the right ($k = 1$ for balloons in Dubray & Laubrock, 2019).

analysis applications. The detection of comic characters/protagonists and faces is a more challenging task, because their appearance can change drastically from one panel to another. Characters can move or be focused on, so they appear in different perspectives and sizes, frontal and profile views, and so on. However, characters also need to be identifiable by the reader, which places constraints on artistic freedom. For instance, they often contain stereotypical elements, such as color patterns (think of Superman's cloak or Obelix's pants).

Very few studies have tried to detect and identify objects, characters, and body parts such as faces in comics. As above, recent progress in computer vision based on deep

CNNs has considerably increased the feasibility of such analyses. Earlier approaches to character detection and identification used SIFT descriptors with local feature matching (Sun, Burie, Ogier, & Kise, 2013) or EOH features after screentone removal, combined with approximate nearest neighbor search (Matsui et al., 2017). Both methods perform well above chance. Since the EOH-based method was developed in the context of a sketch-based image retrieval system, query images used in the evaluation were based on sketches rather than on other pages from the same data set; hence, direct quantitative comparison to other methods is difficult.

Qin, Zhou, He, Wang, and Tang (2017) and Chu and Li (2017) used deep neural networks for face detection in comics. Qin et al. (2017) based their model on Faster R-CNN and achieved very good performance on a manga test set resembling their training material, and fair transfer to the eBDthèque. They note that "compared with Japanese comics, it is more challenging to detect character faces in American-European comics due to the more diverse facial patterns" (p. 1079). Chu and Li (2017) used a more custom-built neural network approach, which computes bounding boxes and character identity from CNN features extracted from regions found by a selective search scheme based on classical features. Both methods outperform classical approaches by a large margin. Possibly, the somewhat better performance of the Qin et al. (2017) model is due to the integrated region proposal in Faster R-CNN, which serves as an attention network.

The multi-task learning models developed by Nguyen et al. (2019) and Ogawa et al. (2018) have been evaluated for character detection on the eBDthèque, for which Nguyen et al. report decidedly better performance. Ogawa et al. (2018) achieved much better performance on the Manga109 data set, which was more similar to their training set. The performance discrepancy between the very similar approaches by Qin et al. (2017) and Nguyen et al. (2019) could also be due to overfitting to differences in the training set. Again this attests to the stylistic differences between Japanese and Western comics and alerts us to the fact that a model should be trained on diverse material if generalizable predictions are desired. To summarize, character detection is a harder problem than panel or speech balloon segmentation, but significant progress has been made. A diverse training set seems even more essential than with other detection tasks.

### *5.7. Visual stylometry*

Comics artists may differ in the drawing style of each of the elements above, and they certainly have their own way of using other stylistic elements, including lines, shapes, colors, textures, or hatching, as well as composition, rhythm, balance, and proportion. Just as writers can be characterized by an analysis of stylometric features such as articles or prepositions (Juola, 2008, 2014), illustrators also have a characteristic visual signature.

Sun and Kise (2011) used HOG feature similarity to successfully identify partial copies of manga and anime faces by other illustrators. Such plagiarism detection can be considered an application of stylometry as it detects unauthorized reuse of style. Dunst and Hartel (2018) show that classical features such as brightness, Shannon entropy, number of shapes, color layout, and edge histograms can be used to differentiate between genres or authorship categories above chance. Chu and Cheng (2016) designed Manga-specific stylistic features based on line, screentones, and panel layout and a style model based on Latent Dirichlet Allocation (LDA), which led to a mean average precision of about 80% in illustrator classification.

Laubrock and Dubray (2018, 2019) were able to correctly predict authorship of pages of both the GNC and the Manga109 data sets with over 92% accuracy based on a combination of deep CNN features. They found that mid-level features corresponding to texture and hatching were more important for classifying illustrator style than higher-level features that are closer to object-level descriptions. As expected, CNN features outperformed classical visual features by a large margin.

Dunst and Hartel (2019a, 2019b) combined classical visual features (see above) and textual features such as word length, standardized type-token ratio, and number of words per page for stylistic classification of graphic novels. They found that the combination of text and image achieved slightly better results with author attribution compared to image data alone, whereas genre classification did not improve. However, such a multimodal stylistic analysis proved successful in distinguishing more or less complex comic books.

Given the results just reported, it is likely that classification accuracy would increase when combined with classical stylometric methods based on text analysis. This requires text extraction as well as computing linguistic features.

## 6. Analysis of text and narrative structure

### *6.1. Text recognition*

As with comics images, most research on text to date has focused on content (or document) analysis. Nevertheless, there are also examples of content generation, as in Saito and Nakamura's (2019) prototype of automated font design for manga. For many contemporary document types, automated text extraction has become routine. However, comics text is characterized by high variability, both in terms of fonts used, text types, and their placement on the comics page. Most text to be found in comics is either hand-drawn by the author or a specialized letterer. Alternatively, fonts are digitized on the basis of an artist's handwriting, or chosen from existing fonts as part of the production process. This diversity complicates computational recognition: the different angles, shapes, and sizes of letters, the varying thickness of hand-drawn lines, and the likelihood of overlapping symbols all lead to higher error rates when applying OCR to comics. This variability finds an equivalent in the different text types that comics employ. Speech representation in the form of monologues or dialogues between characters may be the most common, but contemporary Western comics also contain an increasing amount of narrative text. Japanese manga and related forms frequently use onomatopoeia. Diegetic text, in turn, may be scattered throughout a comics page, from store signs to characters readings newspapers. Graphic boundaries, which can help with balloon or caption spotting, do not enclose diegetic text or onomatopoeia.

There are, of course, established conventions for the placement of text in comics, but many comics artists also decide to go against convention and eschew bubbles or captions. This brief discussion already shows that comics text will usually be distinct from one book or series to another but may also vary within a single narrative, such as in a graphic novel that contains different handwritten fonts for dialogue, narrative and diegetic text. The combination of this variability, the small number of available data sets, and the frequent occurrence of paper degradation or low scan quality means that text extraction and recognition in comics remains an on-going research problem.

The urgent task of high-quality text recognition for comics has usually taken one of two established routes: top-down approaches begin by identifying captions or speech balloons, using one of the approaches described above (Ranjini & Sundaresan, 2013). As a consequence, diegetic text or onomatopoeia are usually disregarded. In contrast, bottom-up approaches seek to detect text by separating the former from graphics without prior layout analysis (Rigaud, Burie, & Ogier, 2017a). Both approaches share their main weakness, namely the reliance on heuristics or generalization for a cultural phenomenon frequently characterized by artistic experimentation and thus deviation from established conventions.

More recently, neural networks have also led to significant advances in recognizing handwriting, rare or irregular fonts. Recent versions of the open-source OCR engines Tesseract (Smith, 2007) and OCRopus (Breuel, 2008) are both based on LSTM neural networks and perform well on historical documents without the need for annotation data. However, these pre-trained networks still lead to relatively high error rates when applied directly to comics (Rayar & Uchida, 2019; Rigaud et al., 2017a). Training simple CNNs on relatively small data sets significantly increases recognition, with Dey et al. (2017) achieving an accuracy rate of 97% on hand-drawn symbols. A different approach is taken in Hartel and Dunst (2019). Using Tesseract 4, the authors show that even imperfectly recognized comics texts can be used for simple text analysis based on a bag-of-words model and yield similar results to manually-transcribed data. While such an approach is not suitable for semantic analysis, it may help with stylistic classification (genre, authorship), particularly once text and image data are combined.

Iyyer et al. (2017) report good results from a commercial OCR (Google CloudVision) applied to pre-segmented text regions. Generally, OCR works much better when regions of interest containing text have been identified in a preprocessing step rather than when the whole comic page scan is input, possibly because stylistic elements such as hatching look similar to text and lead to a large number of false positives. Surprisingly, the literature on scene text spotting (for a recent review see Lin et al., 2019) has largely been ignored by computational comics research. Given that good balloon and caption segmentation are now available (see above), progress in automatic recognition of text is to be expected in the near future, which in turn should allow for models addressing semantic text analysis.

### *6.2. Semantic and syntactic analysis of text*

Given what has been said about the challenges of OCR for comics, it will come as little surprise that instances of semantic text analysis are rare. Those that do exist either draw on text that has been transcribed by hand (Unser-Schutz, 2015) or so-called paratexts that do not form part of the narrative itself, as in Walsh's consideration of fan letters that were amenable to existing OCR software (2018).

To our knowledge, the work by Iyyer et al. (2017) is the sole example of automated semantic text analysis in comics to date. Their approach of encoding the text of each speech balloon as its word embedding sum in combination with LSTMs for intra- and interpanel sequences, though suitable for their purpose of answering cloze questions about balloon order, seems rather too simplistic to capture details of the narrative.

Whereas instances of semantic text analysis remain rare for comics, the recent advances in text detection and recognition may soon change this status quo. Several established and emerging methods may then prove transferable. Computational linguistics offers a wide array of robust approaches on the basis of both bag-of-words and sequential models. These have been adapted for the stylometric analysis of literary texts in the widely used R-package *Stylo* (Eder et al., 2013). Latent Dirichlet allocation (LDA) has been used to generate so-called topic models of text corpora that construct themes or motifs of co-occurring words (Blei et al., 2003). However, the resulting topics depend on subjective input by researchers and only produce results with larger corpora, which might mitigate against their application to existing comics data sets.

As in the case of visual data, deep neural network-based models have become the state of the art for computing text encodings. Since text is inherently sequential, architectures that represent sequences are called for. Within the neural network domain, LSTMs (Hochreiter & Schmidhuber, 1997), a gated variant of recurrent neural networks (RNNs) with long-term memory have traditionally been used to represent sequences (e.g., Wu et al., 2016). Though LSTMs are much easier to train than standard RNNs, they are still not very efficient. Alternative approaches for sequence modeling have recently been proposed, including so-called transformer encoders (Vaswani et al., 2017) and temporal convolutional networks (Gehring et al., 2017). Both are much more efficient to train and better at capturing long-distance dependencies than RNN architectures, but due to their recency the ecosystem of tools is less well-developed, and less is known about their potential weaknesses.

Semantic vector representations of text based on CNNs hold potential for comics. Such word embeddings are dense, learned representations of linguistic material, often obtained from training a sequence-honoring deep neural network model on very large data sets like the Wikipedia or the Wall Street Journal corpus. Analogous to the visual features obtained from training on large image data sets, precomputed word embeddings can be re-used when the data set size is small. Semantic vector representations have been applied to text corpora (Mikolov et al., 2013) and individual illustrations (Saito & Matsui, 2015) with good results.

In contrast to directional word embeddings, which read text sequentially from left to right or right to left, models like BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018) learn context bidirectionally, that is, they represent upcoming in addition to preceding context. Whereas ELMo relies on LSTMs, BERT applies the Transformer network architecture. Both approaches achieve significantly improved results on several NLP tasks, including named-entity recognition, question-answering, and next-sentence prediction. In summary, although little to no work has been done on the syntactic and semantic analysis of comics text, methods developed in other domains may soon be applicable.

### *6.3. Combining image and text*

Given the complex entwining of text and images in comics, a combination of visual and textual topics seems called for. For visual topics, the CNN-based stylometric methods referenced above seem promising, whereas the word embeddings just reviewed might be used for representing text. Additionally, different text types such as dialogue and narrative text should probably be distinguished. These representations can be computed and combined panel-wise (e.g., by concatenation), but a higher-order, hierarchical sequence representation is needed to fully represent narrative development across panels (see below).

In recent years, neural network models for automatic image captioning have seen a surge of interest. These are typically encoder–decoder architectures combining a CNN encoder with some sequence-representing model such as an RNN as a decoder (e.g., Johnson et al, 2016). Such models may have some potential for application in comics analysis. However, one interesting aspect of comics is that text and image are not necessarily directly related. McCloud (1993) lists seven distinct categories of word-image relations. For instance, word and picture may essentially convey the same message, pictures may just illustrate text that could stand alone, words may just add a soundtrack to a picture, and there may even be sections where text and image tell independent stories (see also Cohn, 2016). McCloud (1993, p. 155) observes that "perhaps the most common type of word-picture combination is the interdependent, where words and pictures go hand in hand to convey an idea that neither could convey alone." Thus, it might be that arriving at a plot-level understanding of comics is a decidedly harder problem than captioning of photographs, as somewhat opaque references would need to be tracked by an automated system.

### *6.4. Analysis of narrative structure*

Comics have been characterized as sequential art (Eisner, 1985; McCloud, 1993). So arguably, sequential aspects ranging from panel transitions to narrative structure are what makes comics uniquely interesting. How can narrative sequences be characterized?

Early approaches focused mainly on the transitions between subsequent panels, for which McCloud (1993; see also Saraceni, 2016) developed a taxonomy based on closure —loosely the amount of mental effort required by the reader to connect two subsequent panels. For example, simple zooming in requires fewer inferences than a scene-to-scene transition that involves a change of location and referents.

However, a narrative is more than just a succession of panels: Generally, it involves a sequence of events within a context, a set of characters, and a plot. Furthermore, narratives can also have a hierarchical structure.

Tseng et al. (2018) provide an example of how to systematically track recurring elements, such as people, things, and places across panel frames. Reoccurrence, repetition, and modification of elements are important in constructing generalized event structures. By tracking these identified elements across the narrative and binding together information about them, cohesive chains are built. Such cohesive ties between (re-)appearances guide the reader's interpretation. This approach draws on Zacks and Tversky's (2001) analysis of event segmentation, which deals with the basic units of events and actions, and how people use event structure in perception, understanding, planning, and action.

Neil Cohn has developed a theoretical analysis derived from linguistics and arrived at what is arguably the most sophisticated and coherent theory of visual narrative to date, "Visual Narrative Grammar" (VNG; Cohn, 2013b, 2013c; for a short summary see Cohn, 2018). He argues that "the semantic information in panels maps to categorical roles within a constituent structure, similar to how words play grammatical roles within constituents in syntax" (Cohn, 2018, p. 310).

In VNG, there are five core categories such as *Establisher* or *Peak*. Together, these categories form coherent pieces termed phases of constituency, with only *Peak* being mandatory. Phases in turn belong to narrative arcs. At the core of Cohn's theory is the observation that there is a canonical structure, and linear order, of categories within a phase (Cohn, 2013c, p. 421). Rather complex narrative arc structures can be created by recursion.

Empirical support is available from studies employing event-related potentials (Cohn et al., 2014) and eye movements (Foulsham et al., 2016). Cohn (2013a, 2013c) also provides diagnostic tools for classifying narrative categories as well as tests for constituent boundaries. Thus, Cohn's theoretical model, which separates narrative structures (presentation) from semantic structures (meaning) appears quite ready for implementation. However, Bateman and Wildfeuer (2014) have criticized a grammar-based approach and instead suggested an approach based on dynamic discourse analysis and discourse pragmatics. Possibly, the two approaches could be used as complementary levels of analysis.

What progress have computational approaches made towards automatic analysis of narrative structure? Document analysis and pattern recognition, as described above, certainly have purposes in their own right. Yet, in the context of narrative analysis, they provide only the first building blocks: a formalized, quantitative description of the artwork in a "useful" feature space. There are to our knowledge only a handful of papers on the computational analysis of sequential and narrative structure in comics.

### *6.4.1. Four-scene comics*

In collaboration with five professional comics artists, Ueno (2019) created a data set of 100 four-scene comics (a format akin to the newspaper comic strip), focusing on story patterns. For two common story patterns ("general" and "sudden" with the punch line in the fourth and first panel, respectively), 10 different plots (including specification of location and character) were constructed. Each artist created his own four-scene comic version for each combination of plot and story pattern with his own specific *touch* (a stylistic category). Panels are annotated with respect to the location of character, faces, speech balloons, and the emotional expression of characters. Ueno (2019) performed a first computational analysis using a simple CNN (AlexNet; Krizhevsky et al., 2012), suggesting that specific features in the first scene can discriminate between the "general" and "sudden" patterns, independent of artist. The potential for computational architectures that explicitly represent sequences such as LSTM has not yet been explored with this promising data set.

#### *6.4.2. Amazing mysteries*

An example of such a sequential analysis is given by Iyyer et al. (2017), who had neural network models predict what the text or image of a panel might be, given context from the two preceding panels. They designed several cloze tasks, of which we describe text cloze and visual cloze. In text cloze, the task was to choose which of three candidate texts would fit a text box in the current panel, given the current panel image (and the two context panels). In visual cloze, the candidates consisted of three images (excluding the text) rather than text.

The model combines a text representation and a visual representation in a hierarchical LSTM architecture (Hochreiter & Schmidhuber, 1997). Text in each text box is represented by a word embedding sum, and the sequence of text boxes within a panel is modeled by an intrapanel LSTM. Visual context is represented by activations in one of the upper layers of a pretrained VGG-16 in response to the context panel. These panel-level text (sequence) and image representations are concatenated and fed into an interpanel LSTM, the final hidden state of which is taken as context representation.

The model performs well above chance but significantly lower than a human baseline in both tasks. When text-only, image-only, and image-text versions of the model were compared, as expected, text representations were particularly important for text cloze, and image representations for image cloze. Interestingly, however, providing images in addition to text helped on the text cloze task, whereas providing text in addition to images did not improve performance (or was even detrimental) in the image cloze task, suggesting that it is dominated by visual similarity.

The model developed by Iyyer et al. (2017) is to our knowledge the most sophisticated computational model of sequential dependencies and content in comics books. Compared to vision tasks, in which the current generation of computational models often achieve superhuman performance, classification of sequential content seems much harder, and humans outperform the model by quite a large margin.

To summarize, much computational work remains to be done in capturing the sequential and narrative aspects of comics. In cognitive science, there exist more precise measurements and arguably more agreed-upon models of lower-level aspects of cognition like motion perception (e.g., Hassenstein & Reichardt, 1956) than of higher-level aspects like working memory (e.g., Oberauer & Kliegl, 2006), let alone narrative understanding. As with cognition, it may be natural that the higher-level aspects of comics are more difficult to measure, quantify, and model computationally. Given that document analysis now has reached a state of considerable sophistication, we are excited to see this frontier being tackled soon. Inspiration may come from models of computational narratology (e.g., Finlayson, 2009, 2016).

## 7. Reading of comics

### 7.1. Visual attention

Tracking a user's gaze in real-time provides a direct and relatively unobtrusive measure of overt attention (Kliegl & Laubrock, 2017), which is an indicator of relevance for perception and cognition. Selective attention as a critical filtering mechanism in human information processing strongly reduces computational complexity (e.g., Tsotsos, 2017).

Itti and Koch (2001) describe a classic framework for modeling attention in visual scenes, drawing on feature integration theory (Treisman & Gelade, 1980) and their implementation of bottom-up models of attention (Itti & Koch, 2000; Itti, Koch, & Niebur, 1998). Humans select objects in a scene using both bottom-up saliency and top-down task-dependent cues. Bottom-up saliency is computed from gradients in several pre-attentive topographic feature maps (sensitive to color, orientation, intensity, etc.), which are extracted in parallel at several spatial scales. Mechanisms of center-surround inhibition and long-range spatial suppression lead to competition for saliency within each feature map. After competition, the feature maps are combined into a single topographically organized saliency map. The point of highest saliency is selected, and subsequently suppressed by inhibition of return to prevent attention from getting stuck at a location. Top-down attention can modulate most stages, for example, by differentially weighting feature maps. This feedback mechanism is critically different from the feedforward processing in most of the neural network models described in the earlier parts of this review. Nevertheless, hierarchical CNN models are certainly conceivable as feature extractors in the Itti & Koch framework, and their use might shed new light on topics such as object-based attention and early vs. late selection (Duncan, 1980).

With respect to the reading of comics, finding out what is selected and when can help answer questions about the relative priority and interdependence of image and text for comprehension. Assuming that attention selects informative regions, gaze measurement may also assist in applications such as gaze-contingent presentation or an enhanced annotation tool.

### 7.2. Eye movement studies

A few reception studies have looked at how readers read comics. In one of the earliest experimental studies, Omori et al. (2004) report that panels are skipped more often when the subsequent panel contains prominent speech balloons, indicating that the reader previews the visual structure of the upcoming panel. Foulsham et al. (2016) found that destroying narratively coherent sequential order in six-panel strip caused more regressions and longer viewing times. While this was to be expected, interestingly, they also observed differences in where people looked depending on whether panels were presented in original or random order, which suggests that selection of fixation location is to some degree guided by high-level processes such as narrative understanding.

The largest eye movement corpus to date has been collected by Laubrock et al. (2018). They found that many of the cognitive and oculomotor processes involved in reading and scene perception are also at work during comics reading. Gaze is quite selectively concentrated on text regions and informative image regions such as characters; in contrast, the background receives very few and broadly distributed fixations. Refixations are centered on text, whereas characters are often fixated just once, and have a higher chance to receive first fixations. Selectivity of first fixations as well as the pattern of fixation durations suggests that part of the upcoming panel is pre-processed in parafoveal vision. Kirtley et al. (2018) observed a similar viewing pattern with respect to text and image. Laubrock et al. (2018) additionally report that the CNN-based saliency model Deep Gaze II (Kümmerer et al., 2017) predicts the distribution of empirical fixation locations very well, although the model had only been trained on photographs. This finding suggests that attention is guided by similar features during viewing of comics and real-world scenes.

Tseng et al. (2018) found that eye tracking can also be used to track narrative understanding in comics. They edited parts of the visual scene to remove or modify cues to cohesive event structures and found that these regions received less attention when they were not relevant for the narrative. Similarly, Hutson et al. (2018) found that viewers' event models in working memory affect visual attentional selection while viewing visual narratives. They manipulated the presence of highly inferable actions embedded in picture stories and found increased viewing times for the immediately following image, which were due to an increased number of fixations used for searching inference-relevant information.

Jain et al. (2012) compared the similarity of different viewers' scanpaths between comics, amateur snapshots, and robot pictures, finding the most coherent viewing behavior for comics. They conclude that artists are successful in guiding their readers' attention. We would like to point out that increased scanpath similarity also points to the fact that the abstracted visual scenes represented by line drawings in comics are relatively simple compared to real-world visual scenes. They might therefore be easier to describe, analyze, and control than complex natural scenes in photographs, and hence should be considered more often in the study of scene perception.

We end this section by describing two systems that combine eye tracking data and computational modeling or analysis to arrive at novel applications. Cao, Lau, and Chan (2014) propose a sophisticated system for automatic composition of manga elements. Assuming that the artist carefully chooses how to guide the reader's attention across a page, they construct a probabilistic graphical model that connects the artist's guiding path (as a latent variable) with page layout, panel elements, and viewer attention. The model learns the guiding path from a relatively small set of manga pages with annotations, including panel properties (shot type, motion state, geometric style, center location), characters and balloons (center locations and radii), and eye tracking data about readers' transitions between elements. The trained model can be used to propose a page layout as well as placement of individual elements such as balloons and characters, given user-specified input about the number of panels, input elements (characters and dialogue), and semantics (shot type and motion state). As the method enables automatic creation of attention-directing compositions, it might also have potential for understanding attention itself.

Thirunarayanan et al. (2017) describe an interesting application scenario combining eye tracking and computational methods. They use a modern clustering technique to automatically determine the number of clusters of recorded gaze data, which can then be used to guide object segmentation. A selection of a small number of "interesting" objects could be an important first step in augmentation of comics with digital effects such as animations and stereoscopy. for hand-held devices. As we have seen above, CNN features can be used for fixation prediction as well as semantic segmentation. It will be interesting to see whether a combination of the clustering and CNN-based methods could be used for automatic selection and segmentation of interesting objects, given a database of comic readers' eye movements.

In summary, eye tracking recordings provide a fascinating insight into reader's navigation and reading of comics pages, and they might be useful to guide computational analyses. Some computational models also provide rather good approximations of empirical saliency as measured by eye movements (e.g., Kümmerer et al., 2017).

## 8. Conclusion

We have reviewed computational approaches to comics, which are mostly focused on document analysis. While comics analysis is certainly a niche, it can benefit from developments in related fields. Following recent developments in deep learning techniques, the state of the art has advanced to a level at which rich feature descriptors are available. Most of the simpler document segmentation and recognition tasks now seem solvable, which lays the ground for modeling more complex aspects such as sequential features, image-text relations, and narrative construction. Models of these aspects are needed to arrive at a deeper understanding of a fundamentally sequential art form.

More generally, the availability of modern feature descriptors based on deep learning techniques has the potential to provide insight into perceptual and cognitive processes involving complex material such as comics or visual scenes. Even though deep neural network models are sometimes described as black boxes, their complexity is many orders of magnitudes smaller than that of the human brain and they should therefore be much easier to understand. Furthermore, measurement and manipulation of the activations of single cells and neural assemblies as well as ablation studies can be performed in silico without major ethical concerns. Thus, there is certainly enormous and underexplored potential for genuine insight.

Of course, it has to be kept in mind that the models were often constructed to solve engineering problems rather than for biological plausibility. Adversarial attacks show that humans and computational models might differ rather substantially in what they base their classification decisions on (Goodfellow et al., 2015). Nevertheless, visual processing hierarchies of models based on deep convolutional neural networks have been shown to correspond to processing in the primate ventral visual stream (Bashivan et al., 2019;

Kriegeskorte, 2015; Yamins & DiCarlo, 2016) and are certainly the best available models of mid- and high-level vision to date. We are looking forward to seeing computational models of cognition and attention that relate deep feature descriptions to human behavior and its neural correlates. We also hope that our review has convinced readers from the cognitive and computational sciences that comics are an exciting field of study that bridges scene perception, reading, and even (the ninth) art.

## Acknowledgments

This research was funded by the German Federal Ministry for Science and Education (BMBF), grants no. 01UG1507B to JL and 01UG1507A to AD. We thank Neil Cohn, Nanne van Noord, David Wilmot, and an anonymous reviewer for helpful comments on an earlier version.

## Notes

1. However, work on event structure and narrative structure in general language processing has seen a small renaissance, and possibly results can be extended to the domain of comics.
2. Some recent large-scale datasets of illustrations (BAM!, Wilber et al., 2017) also contain material and annotations related to comics (such as emotion or object labels).
3. A real comics museum, the world-famous Cité international de la bande dessinée in Angoulème, has also digitized part of their large collection and allows online viewing. However, the downloadable page archives are not currently useful for research due to their low resolution, watermarked images.

## References

Arai, K., & Tolle, H. (2011). Method for real time text extraction of digital manga comic. *International Journal of Image Processing (IJIP)*, 669–676.

Augereau, O., Iwata, M., & Kise, K. (2017). An overview of comics research in computer science. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (vol. *3*, pp. 54–59). Kyoto, Japan: Institute of Electrical and Electronics Engineers ( IEEE )/Curran Associates. https://doi.org/10.1109/ICDAR.2017.292

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*, eaav9436. https://doi.org/10.1126/science.aav9436.

Bateman, J. A., Veloso, F. O. D., Wildfeuer, J., Cheung, F. H., & Guo, N. S. (2017). An open multilevel classification scheme for the visual layout of comics and graphic novels: Motivation and design. *Digital Scholarship in the Humanities*, *32*, 476–510. https://doi.org/10.1093/llc/fqw024

Bateman, J. A., & Wildfeuer, J. (2014). A multimodal discourse theory of visual narrative. *Journal of Pragmatics*, *74*, 180–208. https://doi.org/10.1016/j.pragma.2014.10.001

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocaction. *Journal of Machine Learning Research*, *2*, 993–1022.

Breuel, T. M. (2008). The OCRopus open source OCR system. In Proc. SPIE 6815, Document Recognition and Retrieval XV, 68159F–15. https://doi.org/10.1117/12.783598

Cao, Y., Chan, A. B., & Lau, R. W. H. (2012). Automatic stylistic manga layout. *ACM Transactions on Graphics*, *31*(6), 141:1–141:10. https://doi.org/10.1145/2366145.2366160

Cao, Y., Lau, R. W. H., & Chan, A. B. (2014). Look over here: Attention-directing composition of manga elements. *ACM Transactions on Graphics*, *33*(4), 94:1–94:11. https://doi.org/10.1145/2601097.2601183

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*, 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chu, W., & Cheng, W. (2016). Manga-specific features and latent style model for manga style analysis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1332–1336). Shanghai, China: Organized by theInstitute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/ICASSP.2016.7471893

Chu, W.-T., & Li, W.-W. (2017). Manga facenet: Face detection in manga based on deep neural network. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR '17* (pp. 412–415). New York: ACM. https://doi.org/10.1145/3078971.3079031

Cohn, N. (2010). The limits of time and transitions: Challenges to theories of sequential image comprehension. *Studies in Comics*, *1*, 127–147. https://doi.org/10.1386/stic.1.1.127/1

Cohn, N. (2013a). Navigating comics: An empirical and theoretical approach to strategies of reading comic page layouts. *Frontiers in Psychology*, *4*, 186. https://doi.org/10.3389/fpsyg.2013.00186.

Cohn, N. (2013b). *The visual language of comics: Introduction to the structure and cognition of sequential images*. London: Bloomsbury.

Cohn, N. (2013c). Visual narrative structure. *Cognitive Science*, *37*, 413–452. https://doi.org/10.1111/cogs.12016

Cohn, N. (2014). The architecture of visual narrative comprehension: The interaction of narrative structure and page layout in understanding comics. *Frontiers in Psychology*, *5*(680), 1–9. https://doi.org/10.3389/fpsyg.2014.00680

Cohn, N. (2016). A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, *146*, 304–323. https://doi.org/10.1016/j.cognition.2015.10.007

Cohn, N. (2018). Visual language theory and the scientific study of comics. In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.), *Empirical comics research: Digital, multimodal, and cognitive methods* (pp. 305–328). New York, NY: Routledge. https://doi.org/10.4324/9781315185354

Cohn, N. (2019). Structural complexity in visual narratives: Theory, brains, and cross-cultural diversity. In M. Grishakova & M. Poulaki (Eds.), *Narrative complexity and media: Experiential and cognitive interfaces* (pp. 174–199). Lincoln: University of Nebraska Press. https://doi.org/10.2307/j.ctvhktjh6.13

Cohn, N., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2014). The grammar of visual narrative: Neural evidence for constituent structure in sequential image comprehension. *Neuropsychologia*, *64*, 63–70. https://doi.org/10.1016/j.neuropsychologia.2014.09.018

Cohn, N., & Magliano, J. P. (this volume). Why study visual narratives? A framework for studying visual narratives in the cognitive sciences. *Topics in Cognitive Science*.

Cohn, N., Pederson, K., & Taylor, R. (2017). A picture is worth more words over time: Multimodality and narrative structure across eight decades of American superhero comics. *Multimodal Communication*, *6*, 19–37. https://doi.org/10.1515/mc-2017-0003

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In C. Schmid, S. Schmid & M. Poulaki (Eds.), *IEEE Computer Society Conference on Computer Vision and Pattern*

*Recognition (CVPR 05)* (pp. 886–893). Los Alamitos, CA: IEEE Computer Society. https://doi.org/10.1109/CVPR.2005.177

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248–255). Miami, FL: IEEE. https://doi.org/10.1109/CVPRW.2009.5206848

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arxiv*:1810.04805, 1–14.

Dey, S., Dutta, A., Lladós, J., Fornés, A., & Umapada, P. (2017). Shallow neural network model for hand-drawn symbol recognition in multi-writer scenario. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (vol. *3*, pp. 31–32). Kyoto, Japan: IEEE. https://doi.org/10.1109/ICDAR.2017.263

Dubray, D., & Laubrock, J. (2019). Deep CNN-based speech balloon detection and segmentation for comic books. In *2019 15th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1237–1243). Syndey, Australia. https://doi.org/10.1109/ICDAR.2019.00200 (Preprint available as arXiv:1902.08137, 1–10).

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review*, *87*, 272–300. https://doi.org/10.1037/0033-295X.87.3.272

Dunst, A., & Hartel, R. (2018).The quantitative analysis of comics: Towards a visual stylometry of graphic narrative. In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.), *Empirical comics research: Digital, multimodal, and cognitive methods* (pp. 43–61). New York, NY: Routledge. https://doi.org/10.4324/9781315185354

Dunst, A., & Hartel, R. (2019a). Multimodale Stilometrie: Herausforderungen und Potenzial kombinatorischer Bild- und Textanalysen am Beispiel Comics. DHd 2019: Book of Abstracts, Frankfurt [forthcoming].

Dunst, A., & Hartel, R. (2019b). Quantifying complexity in multimodal media: Alan Moore and the “Density” of the Graphic Novel. DH 2019: Book of Abstracts, Utrecht [forthcoming].

Dunst, A., Hartel, R., & Laubrock, J. (2017). The Graphic Narrative Corpus (GNC): Design, annotation, and analysis for the Digital Humanities. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (vol. *3*, pp. 15–20). Kyoto, Japan: IEEE. https://doi.org/10.1109/ICDAR.2017.286

Dunst, A., J. Laubrock, & J. Wildfeuer (Eds.) (2018). *Empirical comics research: Digital, multimodal, and cognitive methods*. New York: Routledge. https://doi.org/10.4324/9781315185354

Eder, M., Kestemont, M., & Rybicki, J. (2013). Stylometry with R: A suite of tools. In *Digital humanities 2013: Conference abstracts* (pp. 487–489). Lincoln: Alliance of Digital Humanities Organizations (ADHO).

Eisner, W. (1985). *Comics & sequential art*. Tamarac, FL: Poorhouse Press.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*, 777–813. https://doi.org/10.1037/0033-295X.112.4.777

Engbert, R., Trukenbrod, H. A., Barthelmé, S., & Wichmann, F. A. (2015). Spatial statistics and attentional dynamics in scene viewing. *Journal of Vision*, *15*(1), 1–17. https://doi.org/10.1167/15.1.14

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2015). The PASCAL visual object classes challenge: A retrospective. *International Journal of Computer Vision*, *111*, 98–136. https://doi.org/10.1007/s11263-014-0733-5

Finlayson, M. A. (2009). Deriving narrative morphologies via analogical story merging. In. B. Kokinov, K. Holyoak, & D. Gentner (Eds.), *New Frontiers in analogy research (Proceedings of the Second International Conference on Analogy)* (pp. 127–136). Sofia: New Bulgarian University Press.

Finlayson, M. A. (2016). Inferring Propp’s functions from semantically annotated texts. *Journal of American Folklore*, *129*, 53–75.

Forceville, C. (2005). Visual representations of the idealized cognitive model of anger in the Asterix album La Zizanie. *Journal of Pragmatics*, *37*, 69–88. https://doi.org/10.1016/j.pragma.2003.10.002

Forceville, C., Veale, T., & Feyaerts, K. (2010). Balloonics: The visuals of balloons in comics. In J. Goggin & D. Hassler-Forest (Eds.), *The rise and reason of comics and graphic literature: Critical essays on the form* (pp. 56–73). Jefferson, NC: McFarland.

Foulsham, T., Wybrow, D., & Cohn, N. (2016). Reading without words: Eye movements in the comprehension of comic strips. *Applied Cognitive Psychology*, *30*, 566–579. https://doi.org/10.1002/acp.3229

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *arXiv*:1705.03122, 1–15.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203772157

Girshick, R. (2015). Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 1440–1448). https://doi.org/10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580–587). https://doi.org/10.1109/CVPR.2014.81

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015).Explaining and harnessing adversarial examples. In Y. Bengio & Y. LeCun (Eds.) ,*3rd International Conference on Learning Representations, ICLR 2015* (pp. 1–11). arXiv:1412.6572.

Greene, M. R., & Oliva, A. (2009a). The briefest of glances: the time course of natural scene understanding. *Psychological Science*, *20*, 464–472. https://doi.org/10.1111/j.1467-9280.2009.02316.x

Greene, M. R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*, 137–176. https://doi.org/10.1016/j.cogpsych.2008.06.001

Guérin, C., Rigaud, C., Mercier, A., Ammar-Boudjelal, F., Bertet, K., Bouju, A., Burie, J., Louis, G., Ogier, J., & Revel, A. (2013). eBDthèque: A representative database of comics. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 1145–1149). Washington, DC: IEEE Computer Society. https://doi.org/10.1109/ICDAR.2013.232

Hartel, R., & Dunst, A. (2019).How good is good enough? Establishing quality thresholds for the automatic text analysis of retro-digitized comics. In I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W. Cheng, & S. Vrochidis (Eds.), *MultiMedia Modeling. MMM 2019. Lecture Notes in Computer Science* (pp. 662–671). Cham: Springer. https://doi.org/10.1007/978-3-030-05716-9_59

Hassenstein, B., & Reichardt, W. (1956). Systemtheoretische Analyse der Zeit-, Reihenfolgen- und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers Chlorophanus. *Zeitschrift für Naturforschung B*, *11*, 513–524. https://doi.org/10.1515/znb-1956-9-1004

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969). Venice, Italy: IEEE. https://doi.org/10.1109/ICCV.2017.322

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243–271. https://doi.org/10.1146/annurev.psych.50.1.243

Ho, A. K. N., Burie, J.-C., & Ogier, J.-M. (2012). Panel and speech balloon extraction from comic books. In *2012 10th IAPR international workshop on document analysis systems* (pp. 424–428). https://doi.org/10.1109/DAS.2012.66

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hutson, J. P., Magliano, J. P., & Loschky, L. C. (2018). Understanding moment-to-moment processing of visual narratives. *Cognitive Science*, *42*, 2999–3033. https://doi.org/10.1111/cogs.12699

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506. https://doi.org/10.1016/S0042-6989(99)00163-7

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 1254–1259. https://doi.org/10.1109/34.730558

Iyyer, M., Manjunatha, V., Guha, A., Vyas, Y., Boyd-Graber, J. L., III, Daume, H., & Davis, L. S. (2017). The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (pp. 6478–6487). Honolulu, HI. IEEE. https://doi.org/10.1109/CVPR.2017.686

Jain, E., Sheikh, Y., & Hodgins, J. (2012). Inferring artistic intention in comic art through viewer gaze. In *Proceedings of the ACM Symposium on Applied Perception, SAP '12* (pp. 55–62). New York, NY: ACM. https://doi.org/10.1145/2338676.2338688

Johnson, J., Karpathy, A., & Li, F.-F. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4565–4574). Las Vegas, NV: IEEE. https://doi.org/10.1109/CVPR.2016.494

Juola, P. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval*, *1*, 233–334. https://doi.org/10.1561/1500000005

Juola, P. (2014). The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, *30*, i100–i113. https://doi.org/10.1093/llc/fqv040

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Kirtley, C., Murray, C., Vaughan, P. B., & Tatler, B. W. (2018).Reading words and images: Factors influencing eye movements in comic reading. In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.), *Empirical approaches to comics research: Digital, multimodal, and cognitive methods* (pp. 264–283). New York, NY: Routledge. https://doi.org/10.4324/9781315185354

Kliegl, R., & Laubrock, J. (2017). Eye-movement tracking during reading. In A. M. de Groot & P. Hagoort (Eds.), *Research methods in psycholinguistics and the neurobiology of language: A practical guide* (pp. 68–88). Hoboken, NJ: Wiley-Blackwell.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Reviews of Vision Science*, *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (vol. *25*, pp. 1097–1105). La Jolla, CA: Neural Information Processing Systems. https://doi.org/10.1145/3065386

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017). Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 4799–4808). Los Alamitos, CA: IEEE. https://doi.org/10.1109/ICCV.2017.513

Laubrock, J., Cajar, A., & Engbert, R. (2013). Control of fixation duration during scene viewing by interaction of foveal and peripheral processing. *Journal of Vision*, *13*(12), 11–20. https://doi.org/10.1167/13.12.11

Laubrock, J., & Dubray, D. (in press). Multi-class semantic segmentation of comics: A U-Net based approach. Paper presented at Graphics Recognition (GREC) workshop, International Conference on Document Analysis and Recognition (ICDAR), Sydney. Available at https://grec2019.univ-lr.fr/wp-content/uploads/2019/09/Booklet_GREC2019.pdf.

Laubrock, J., & Dubray, D. (2018). Computational analysis and visual stylometry of comics using convolutional neural networks. In J. G. Palau & I. G. Russell (Eds.), *Digital humanities 2018 puentes-bridges* (pp. 228–231). Mexico City: Red de Humanidades Digitales A. C.

Laubrock, J., & Dubray, D. (2019). CNN-based classification of illustrator style in graphic novels: Which features contribute most? In I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W. H. Cheng, & S. Vrochidis (Eds.), *MultiMedia modeling. MMM 2019. Lecture Notes in Computer Science*, vol. *11296*. (pp. 684–695). Cham: Springer. https://doi.org/10.1007/978-3-030-05716-9_61

Laubrock, J., Hohenstein, S., & Kümmerer, M. (2018).Attention to comics: Cognitive processing during the reading of graphic literature. In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.), *Empirical comics research: Digital, multimodal, and cognitive methods* (pp. 239–263). New York: Routledge. https://doi.org/10.4324/9781315185354

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten ZIP code recognition. *Neural Computation*, *1*, 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Levi, K., & Weiss, Y. (2004). Learning object detection from a small number of examples: The importance of good features. In *Proceedings of the CVPR* (vol. *2*, pp. 53–60). Los Alamitos, CA: IEEE. https://doi.org/10.1109/CVPR.2004.1315144

Lin, H., Yang, P., & Zhang, F. (2019). Review of scene text detection and recognition. *Archives of Computational Methods in Engineering*, *27*. https://doi.org/10.1007/s11831-019-09315-1

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 936–944). Los Alamitos, CA: IEEE. https://doi.org/10.1109/CVPR.2017.106

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision—ECCV 2014* (pp. 740–755). Cham: Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016).SSD: Single shot multibox detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision—ECCV 2016* (pp. 21–37). Cham: Springer. https://doi.org/10.1007/978-3-319-46448-0_2

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3431–3440). https://doi.org/10.1109/CVPR.2015.7298965

Loschky, L. C., Hutson, J. P., Smith, M. E., Smith, T. J., & Magliano, J. (2018). Viewing static visual narratives through the lens of the Scene Perception and Event Comprehension Theory (SPECT). In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.), *Empirical comics research: Digital, multimodal, and cognitive methods* (pp. 217–238). New York, NY: Routledge. https://doi.org/10.4324/9781315185354

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*, 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Manovich, L. (2012). How to compare one million images?In D. M. Berry (Ed.) *Understanding digital humanities* (pp. 249–278). London: Palgrave Macmillan. https://doi.org/10.1057/9780230371934_14

Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., & Aizawa, K. (2017). Sketch-based manga retrieval using Manga109 dataset. *Multimedia Tools and Applications*, *76*, 21811–21838. https://doi.org/10.1007/s11042-016-4020-z

McCloud, S. (1993). *Understanding comics*. New York: Harper Perennial.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 3111–3119). Proceedings of the 26th International Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems 26. Lake Tahoe. NV: Neural Information Processing Systems Foundation, Inc. ( NIPS )

Morton, G. M. (1966). *A computer oriented geodetic data base and a new technique in file sequencing*. Ottawa: International Business Machines Co.

Nguyen, N.-V., Rigaud, C., & Burie, J.-C. (2018). Digital comics image indexing based on deep learning. *Journal of Imaging*, *4*(7), 1–34. https://doi.org/10.3390/jimaging4070089

Nguyen, N. V., Rigaud, C., & Burie, J.-C.(2019). Multi-task model for comic book image analysis. In I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W. H. Cheng, & S. Vrochidis (Eds.), *MultiMedia modeling. MMM 2019*. Lecture Notes in Computer Science, vol. *11296*. Cham: Springer. https://doi.org/10.1007/978-3-030-05716-9_57

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*, 601–626. https://doi.org/10.1016/j.jml.2006.08.009

Ogawa, T., Otsubo, A., Narita, R., Matsui, Y., Yamasaki, T., & Aizawa, K. (2018). Object detection for comics using Manga109 annotations. *arXiv*:1803.08670, 1–18.

Ogier, J.-M., Aizawa, K., Osaka, K. K., Burie, J.-C., Yamasaki, T., & Osaka, M. I. (2016).Proceedings of the 1st international workshop on coMics ANalysis, Processing and Understanding. Association for Computing Machinery.

Omori, T., Ishii, T., & Kurata, K. (2004).Eye catchers in comics: Controlling eye movements in reading pictorial and textual media. 28th International Congress of Psychology, Beijing, China.

Pang, X., Cao, Y., Lau, R. W., & Chan, A. B. (2014). A robust panel extraction method for manga. In Proceedings of the 22Nd ACM International Conference on Multimedia MM '14 (pp. 1125-1128). New York: ACM. https://doi.org/10.1145/2647868.2654990

Pederson, K., & Cohn, N. (2016). The changing pages of comics: Page layouts across eight decades of American superhero comics. *Studies in Comics*, *7*, 7–28. https://doi.org/10.1386/stic.7.1.7_1

Peters, M. E., Neumann, M., Iyyer, M., & Gardner, M. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. *1*, pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. 10.18653/v1/N18-1202

Qin, X., Zhou, Y., He, Z., Wang, Y., & Tang, Z. (2017). A faster R-CNN based method for comic characters face detection. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (vol. *01*, pp. 1074–1080). https://doi.org/10.1109/ICDAR.2017.178

Ranjini, S., & Sundaresan, M. (2013). Extraction and recognition of text from digital English comic image using median filter. *International Journal on Computer Science and Engineering (IJCSE)*, *5*, 238–244.

Rayar, F., & Uchida, S. (2019).Comic text detection using neural network approach.In I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W. H. Gurrin (Eds.), *MultiMedia modeling. MMM 2019. Lecture Notes in Computer Science*, vol *11296*. pp. 672–683). Cham: Springer. https://doi.org/10.1007/978-3-030-05716-9_60

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. (2012). *Psychology of reading* (2nd ed.). New York: Taylor & Francis. https://doi.org/10.4324/9780203155158

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863. https://doi.org/10.1016/S0022-5371(67)80149-X

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). You only look once: Unified, real-time object detection. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 779–788). Los Alamitos, CA: IEEE. https://doi.org/10.1109/CVPR.2016.91

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6517–6525). Los Alamitos, CA: IEEE. https://doi.org/10.1109/CVPR.2017.690

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Proceedings of the 28th International Conference on Neural Information Processing Systems – Volume 1, NIPS'15* (pp. 91–99). Cambridge, MA: MIT Press.

Rigaud, C., Burie, J., & Ogier, J. (2017a). Segmentation-free speech text recognition for comic books. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (vol. *3*, pp. 29–34). Los Alamitos, CA: IEEE. https://doi.org/10.1109/ICDAR.2017.288

Rigaud, C., Burie, J.-C., & Ogier, J.-M. (2017). Text-independent speech balloon segmentation for comics and manga. In B. Lamiroy & R. D. Lins (Eds.), *Graphic recognition. Current trends and challenges* (pp. 133–147). Cham: Springer. https://doi.org/10.1007/978-3-319-52159-6_10

Rigaud, C., Burie, J.-C., Ogier, J.-M., Karatzas, D., & van de Weijer, J. (2013). An active contour model for speech balloon detection in comics. In *12th International Conference on Document Analysis and Recognition (ICDAR 2013)* (pp. 1240–1244). Los Alamitos, CA: IEEE. https://doi.org/10.1109/ICDAR.2013.251

Rigaud, C., Guérin, C., Karatzas, D., Burie, J.-C., & Ogier, J.-M. (2015). Knowledge-driven understanding of images in comic books. *International Journal on Document Analysis and Recognition (IJDAR)*, *18*, 199–221. https://doi.org/10.1007/s10032-015-0243-1

Rigaud, C., Thanh, N. L., Burie, J. C., Ogier, J. M., Iwata, M., Imazu, E., & Kise, K. (2015). Speech balloon and speaker association for comics and manga understanding. In *Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 351–355). Los Alamitos, CA: IEEE. https://doi.org/10.1109/ICDAR.2015.7333782

Rigaud, C., Tsopze, N., Burie, J.-C., & Ogier, J.-M.(2013). Robust frame and text extraction from comic books.In Y.-B. Kwon & J.-M. Ogier (Eds.), *Graphics recognition. New Trends and challenges* (pp. 129–138). Berlin: Springer. https://doi.org/10.1007/978-3-642-36824-0_13

Ronneberger, O., Fischer, P., & Brox, T. (2015).U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. Wells, & A. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015*. Lecture Notes in Computer Science (vol. *9351*). Cham: Springer.

Rosenfeld, A., & Pfaltz, J. L. (1966). Sequential operations in digital picture processing. *Journal of the ACM*, *13*, 471–494. https://doi.org/10.1145/321356.321357

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Saito, J., & Nakamura, S. (2019). Fontender: Interactive Japanese Text Design with Dynamic Font Fusion Method for Comics. In I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W. H. Cheng, & S. Vrochidis (Eds.), *MultiMedia Modeling. MMM 2019*. Lecture Notes in Computer Science (vol. *11296*). Cham: Springer.

Saito, M., & Matsui, Y. (2015). Illustration2vec: A semantic vector representation of illustrations. In *SIGGRAPH Asia 2015 Technical Briefs, SA '15* (pp. 5:1–5:4). New York, NY: ACM. https://doi.org/10.1145/2820903.2820907

Saraceni, M. (2016). Relatedness: Aspects of textual connectivity in comics. In N. Cohn (Ed.), *The visual narrative reader* (pp. 115–127). New York: Bloomsbury.

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Engbert, R., & Wichmann, F. A. (2019). Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of Vision*, *19*(3), 1. https://doi.org/10.1167/19.3.1.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Smith, R. (2007). An overview of the Tesseract OCR Engine. In *9th International Conference on Document Analysis and Recognition (ICDAR)* (vol. *2*, pp. 629–633). Los Alamitos, CA: IEEE. https://doi.org/10.1109/ICDAR.2007.4376991

Spiegelman, A. (1997). *Maus: A survivor's tale*. New York, NY: Pantheon Books.

Stommel, M., Merhej, L. I., & Müller, M. G. (2012). Segmentation-free detection of comic panels. In L. Bolc, R. Tadeusiewicz, L. J. Chmielewski, & K. Wojciechowski*Proceedings of the International Conference on Computer Vision and Graphics—Volume 7594, ICCVG 2012* (pp. 633–640). New York, NY: Springer. https://doi.org/10.1007/978-3-642-33564-8_76

Sun, W., Burie, J., Ogier, J., & Kise, K. (2013). Specific comic character detection using local feature matching. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 275–279). https://doi.org/10.1109/icdar.2013.62

Sun, W., & Kise, K. (2011). Similar partial copy recognition for line drawings using concentric multi-region histograms of oriented gradients. In *Proceedings of the IAPR Conference on Machine Vision Applications* (pp. 71–74). Nara, Japan. http://www.mva-org.jp/Proceedings/2011CD/papers/04-09.pdf

Tanaka, T., Shoji, K., Toyama, F., & Miyamichi, J. (2007). Layout analysis of tree-structured scene frames in comic images. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI'07* (pp. 2885–2890). San Francisco: Morgan Kaufmann.

Tatler, B. W., Brockmole, J. R., & Carpenter, R. H. S. (2017). LATEST: A model of saccadic decisions in space and time. *Psychological Review*, *124*, 267–300. https://doi.org/10.1037/rev0000054

Thirunarayanan, I., Khetarpal, K., Koppal, S., Meur, O. L., Shea, J., & Jain, E. (2017). Creating segments and effects on comics by clustering gaze data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *13(3)*, 24:1–24:23. https://doi.org/10.1145/3078836

Treisman, A., & Gelade, G. A. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Tseng, C.-I., Laubrock, J., & Pflaeging, J. (2018).Character developments in comics and graphic novels: A systematic analytical scheme. In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.) (pp. 154–175). New York: Routledge. https://doi.org/10.4324/9781315185354

Tsotsos, J. K. (2017). Complexity level analysis revisited: What can 30 years of hindsight tell us about how the brain might represent visual information? *Frontiers in Psychology*, *8*(1216), 1–16. https://doi.org/10.3389/fpsyg.2017.01216

Ueno, M. (2019).Structure analysis on common plot in four-scene comic story dataset. In I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W. H. Cheng, & S. Vrochidis (Eds.), *MultiMedia Modeling. MMM 2019*. Lecture Notes in Computer Science (vol. *11296*). Cham: Springer.

Unser-Schutz, G. (2015). Influential or influenced? The relationship between genre, gender and language in manga. *Gender & Language*, *9*, 223–254. https://doi.org/10.1558/genl.v9i2.17331

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (vol. *30*, pp. 5998–6008). Long Beach, CA.

Walsh, J. A. (2012). Comic book markup language: An introduction and rationale. *Digital Humanities Quarterly*, 6.

Walsh, J. A., Martin, S., & St. Germain, J. (2018)."The spider's web": An analysis of fan mail from amazing spider-man, 1963–1995. In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.) (pp. 62–84). New York, NY: Routledge. https://doi.org/10.4324/9781315185354

Ware, C. (2012). *Building stories*. New York: Pantheon Books.

Wilber, M. J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J., & Belongie, S. (2017). BAM! The behance artistic media dataset for recognition beyond photography. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 1211–1220). Los Alamitos, CA: IEEE. https://doi.org/10.1109/iccv.2017.136

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*:1609.08144, 1–23.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356. https://doi.org/10.1038/nn.4244

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*, 3–21. https://doi.org/10.1037/0033-2909.127.1.3

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, *23*, 1028–1034. https://doi.org/10.3758/s13423-015-0864-x