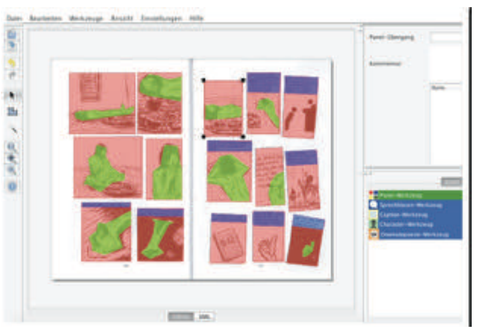
Design of a corpus and an interactive annotation tool for graphic literature

Jochen Laubrock, Eike Richter & Sven Hohenstein

University of Potsdam, Germany – laubrock@uni-potsdam.de

Comics and graphic novels constitute a unique cultural form that has developed its own vocabulary, allowing for a fascinating interplay of text and visual art. After a period of neglect, they have recently been theoretically analyzed in detail by scholars in the arts, humanities and linguistics (McCloud, 1993; Groensteen, 2007; Cohn, 2013). Our aim is to provide an empirical testbed for these theories. To this end, we develop a richly annotated curated corpus of graphic novels.

Subjective Description

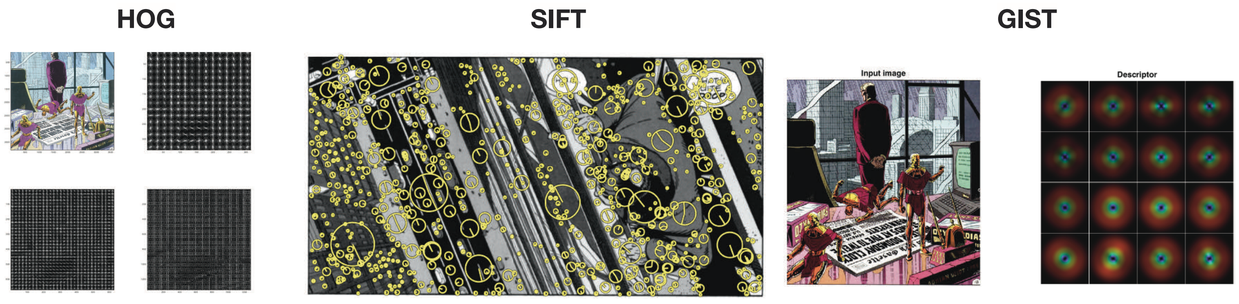


The foundation of this endeavor is a large collection of graphic novels, which are annotated using a variety of methods. Part of the annotations are provided by human coders, and include *high-level* meta data about the works, and *mid-level* descriptions of pages and panels, including the locations and identities of actors / characters, text, objects, and panel transitions. To aid human coders, we developed an annotation tool and a corresponding XML dialect, extending CBML (Walsh, 2012).

Objective Description

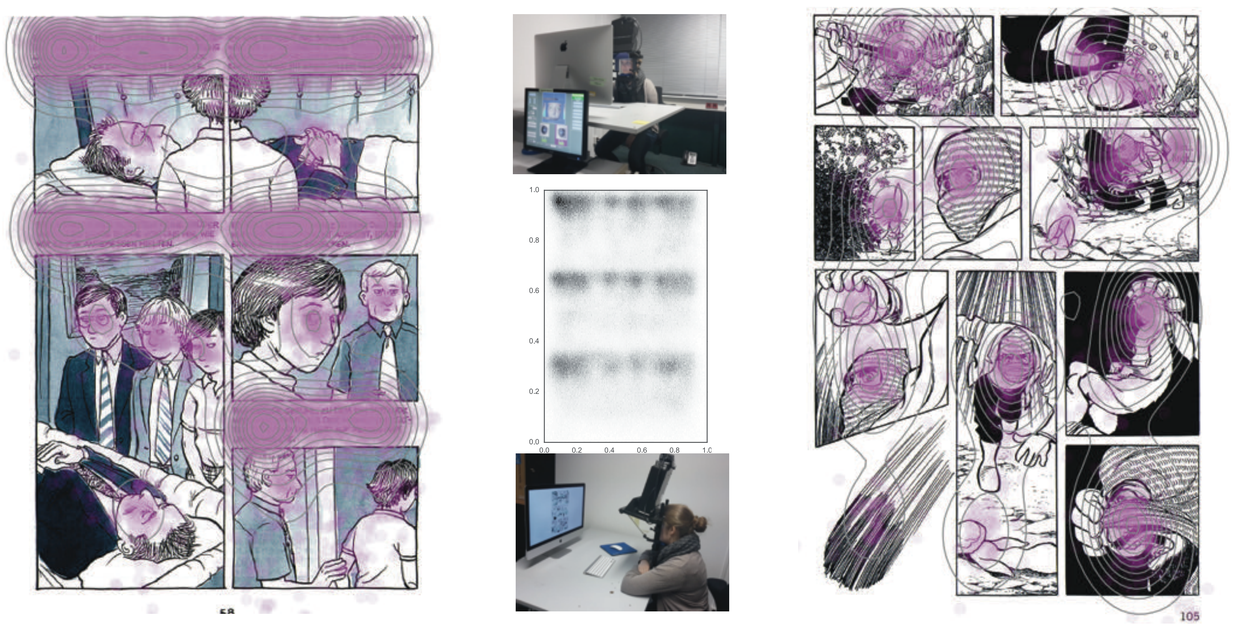
Since human descriptions necessarily provide a degree of subjectiveness, we also include more objective *lower-level* descriptions of visual elements in terms of feature descriptors developed in computer vision, such as SIFT or HOG. We are currently evaluating the addition of *mid-level* features from deep networks trained on photographs of real-world scenes, with quite promising first results.

HOG **SIFT** **GIST**



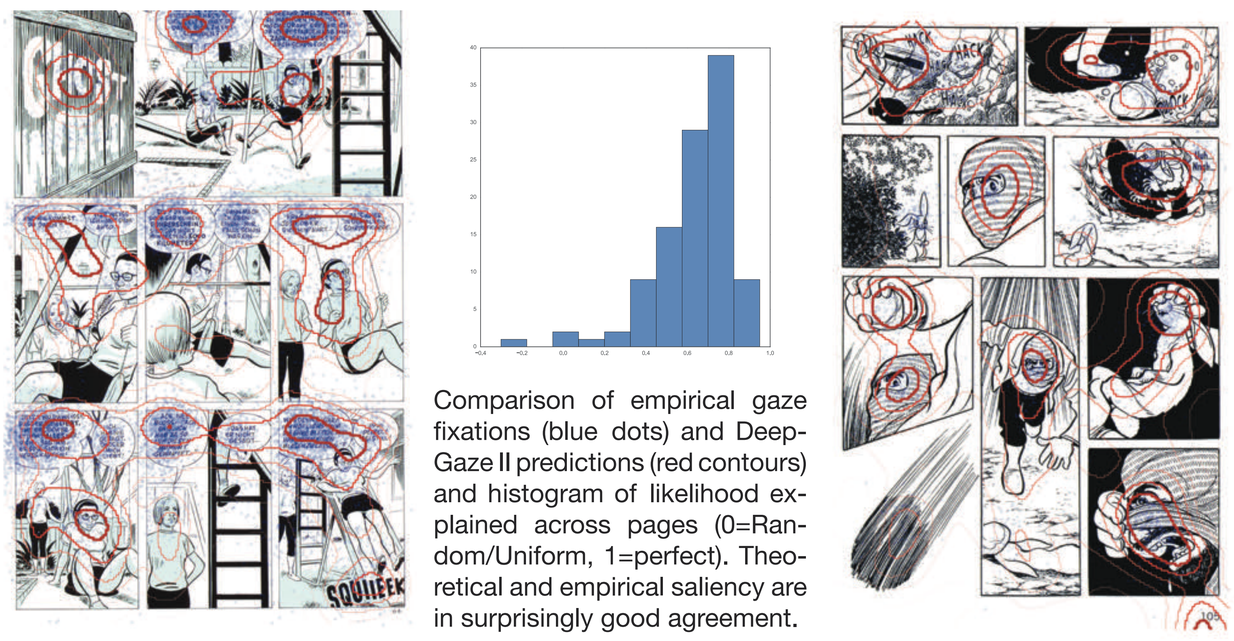
Reader-Level Description

Annotations are augmented by eye-tracking data for a subset of pages, reflecting where readers direct their attention (*empirical saliency*). The corpus can therefore also be used for evaluating computational models of saliency (*theoretical saliency*).



Deep Neural Networks (e.g., VGG-19, Simonyan et al., 2014) trained for object classification using the large ImageNet database (organized along the nouns of the WordNet hierarchy) develop feature detectors similar to those in the human visual system. DeepGaze II (Kümmerer et al., 2016) uses VGG19 feature responses to train a second Neural network to predict empirical saliency, i.e. pseudo-fixations on a new set of images (SALICON, MIT300). DeepGaze-II saliency predictions appear to generalize very well to drawings in comics.

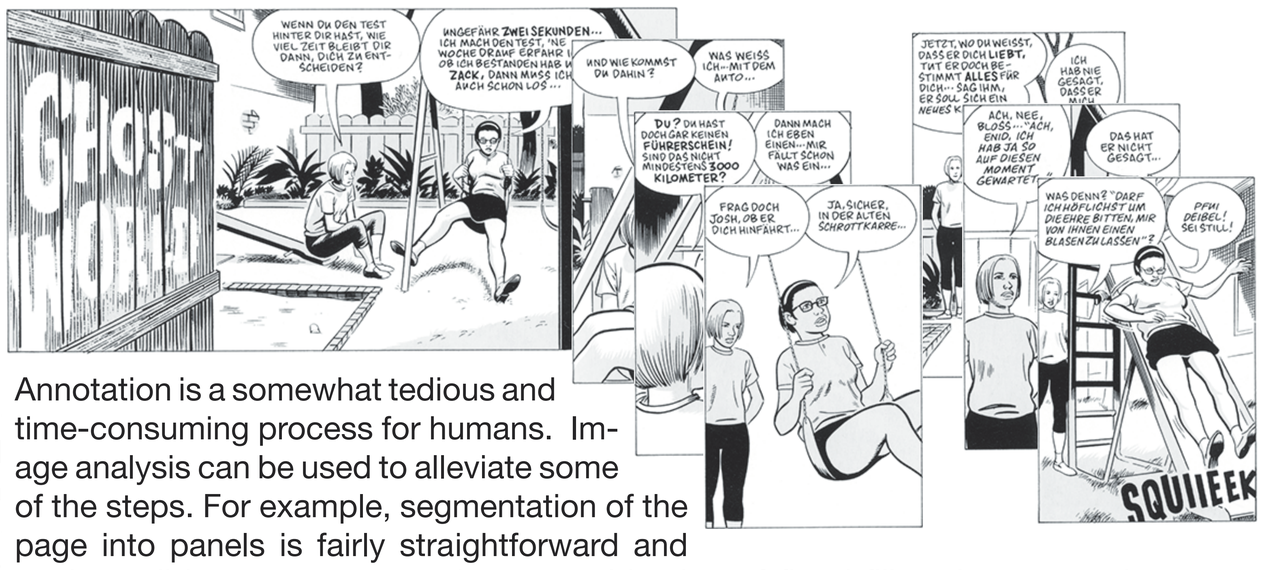
Deep Neural Networks: VGG-19 + DeepGaze II



Comparison of empirical gaze fixations (blue dots) and DeepGaze II predictions (red contours) and histogram of likelihood explained across pages (0=Random/Uniform, 1=perfect). Theoretical and empirical saliency are in surprisingly good agreement.

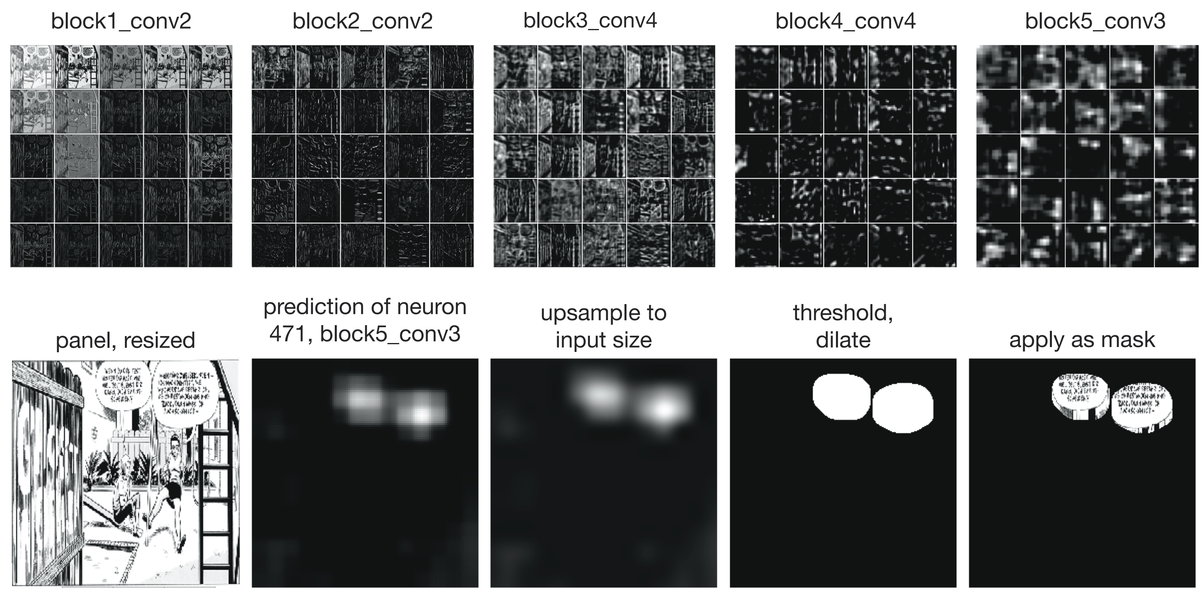
Improving the annotation tool

Panel segmentation



Annotation is a somewhat tedious and time-consuming process for humans. Image analysis can be used to alleviate some of the steps. For example, segmentation of the page into panels is fairly straightforward and works well in many cases, using a combination of flood-fill and connected-components operation, given some initial heuristic guess (or human input) of which is the likely background color, and some post-processing discarding very small regions.

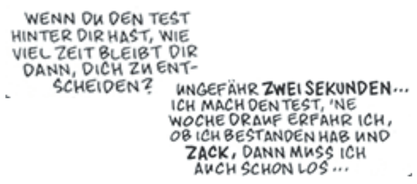
Text spotting



Automatically locating text in a comics panel is surprisingly difficult, given the ease with which humans can perform the task. While classical algorithms are satisfactory for panel detection, they perform less well for recognition of other elements such as text, speech bubbles, captions, or persons. Deep features might be useful for these tasks. For example, some neurons in the upper convolutional layers of pre-trained VGG-19 can be used as a speech bubble detectors. Additional training using gold standard annotations is likely to yet improve the results. Projected further developments include recognition of faces and protagonist identities.

Text recognition

After text regions have been spotted, automated text recognition using optical character recognition (OCR) can be attempted. While OCR is often considered solved, the tiny and hand-lettered fonts pose some problems. In fact, Iiyer (2016) recently reported that OCR on comics text was harder than many other classification tasks. Our preprocessing pipeline involves text spotting (above), up-sampling, thresholding, skew correction, and optional vectorizing using potrace. OCR engines like tesseract still only provide reasonable performance after training on the specific font, stressing the need for initial manual annotation. OCR training will be added to the annotation tool.




```
" WENN OU OEN TEST HINTER DIR HAST, WIE VIE ZEIT BLEIBT
DIR OANN, DICH ZA ENT SCHEIDEN? L UNCEFÁHRZWEISEKUNDEN
... ICH MACHOEN TEST, 'NE WOCHE ORAWF ERPAHR ACH, O8 iCH
BEST ANDEN HAB ND ZACK DANN MnSS LCH AVCH SCHON LOS ... "
```

Applications

An annotated corpus of graphic novels is a valuable resource for *linguistics* and *literature studies*. For example, stylometric analyses of text, but also of visual features becomes possible. The set of tools and methods is potentially useful for other fields, such as *arts history* (cf. Elgammal & Saleh, 2015). Availability of ground truth annotations provides training material for object recognition and classification tasks in *computer science* and *data science*. *Cognitive psychology* might be interested in using the combination of eye movement data, image and text descriptions for evaluating models of attention during reading and scene perception. The AOI-based annotations additionally allow for interesting experimental tests of information integration across panels or narrative construction. More generally, a set of tools based on image processing might benefit various areas of the *Digital Humanities*.

References

Cohn, N. (2013). *The Visual Language of Comics. Introduction to the Structure and Cognition of Sequential Images.* London: Bloomsbury.
Elgammal, A. & Saleh, B. (2015). Quantifying Creativity in Art Networks. *6th International Conference on Computational Creativity* (ICCC'15).
Groensteen, Thierry. 2007. *The System of Comics.* Translated by B. Beaty and N. Nguyen. Jackson, MI: University of Mississippi Press.
Kümmerer, M., Wallis, T.S.A., & Bethge, M.: DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv*:1610.01563
McCloud, S. (1993). *Understanding Comics: The Invisible Art.* New York: Harper Collins.
Simonyan, K. & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556, *arXiv*:1409.1556.
Walsh, J. (2012). Comic Book Markup Language: An Introduction and Rationale. *Digital Humanities Quarterly,* 6(1).

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung