Graph-Based Format for Modeling Multimodal Annotations in Virtual Reality by Means of VAnnotatoR

Giuseppe Abrami(✉), Alexander Mehler, and Christian Spiekermann

Texttechnology Lab, Goethe University Frankfurt, Frankfurt, Germany
abrami@em.uni-frankfurt.de

Abstract. Projects in the field of *Natural Language Processing* (NLP), the *Digital Humanities* (DH) and related disciplines dealing with machine learning of complex relationships between data objects need annotations to obtain sufficiently rich training and test sets. The visualization of such data sets and their underlying *Human Computer Interaction* (HCI) are perennial problems of computer science. However, despite some success stories, the clarity of information presentation and the flexibility of the annotation process may decrease with the complexity of the underlying data objects and their relationships. In order to face this problem, the so-called VAnnotatoR was developed, as a flexible annotation tool using 3D glasses and augmented reality devices, which enables annotation and visualization in three-dimensional virtual environments. In addition, multimodal objects are annotated and visualized within a graph-based approach.

Keywords: Annotation · Virtual Reality · VAnnotatoR · Image analysis · Digital Humanities

1 Introduction

Projects in the field of *Natural Language Processing* (NLP), the *Digital Humanities* (DH) and related disciplines dealing with machine learning of complex relationships between data objects need annotations to obtain sufficiently rich training and test sets. The visualization of such data sets and their underlying *Human Computer Interaction* (HCI) are perennial problems of computer science. In any event, visualizing annotations is an old topic, and various projects (e.g. [10,11,17,22]) have experimented with mostly 2D models thereby achieving considerable success. Despite these success stories, one can nevertheless observe that the clarity of information presentation and the flexibility of the annotation process decrease with the complexity of the underlying data objects and their relations. In order to overcome these pitfalls, the interface of annotation tools needs to be developed considerably: It must leave behind the narrow tracks of 2D graph-like visualizations that make annotations confusing as soon as a certain

C. Stephanidis and M. Antona (Eds.): HCII 2019, CCIS 1088, pp. 351–358, 2019.
https://doi.org/10.1007/978-3-030-30712-7_44

number of annotation units and their relations is exceeded. And it must become much more interactive by ideally exploiting the full bandwidth of human multimodal information processing to map objects and their network relations. In recent years, technological progress in HCI has largely focused on the development of 3D interfaces, if performance and flexibility optimization is ignored. However, the implementation of 3D interfaces in the field of manual annotations of information objects is still pending. Therefore, one can state that interfaces of current annotation tools are still a barrier for new forms of visualization and interaction with information objects.

In order to overcome these barrier and to implement intuitive annotation methods, the so-called VAnnotatoR [20] was developed. It uses state-of-the-art technologies of *Virtual Reality* (VR) and *Augmented Reality* (AR) to address the interface problem. VAnnotatoR (VAR) is not the only annotator or visualizer in VR (e.g. [6,21]) but differs from other projects in terms of its flexible data model and the collaborative parallel use that it enables in AR.

This paper will exemplify the annotation of instances of the graph-like structures (as enumerated by Fig. 1) by means of VAR. Furthermore, it will describe the capabilities of VAR regarding the implicit, inferential annotation of object relations. This will be done by example of text-image and text-image-building relations where the buildings are animated as walk-on-able artifacts.

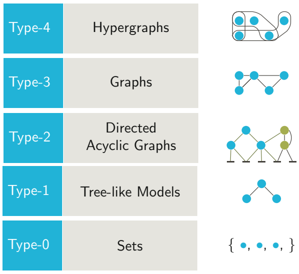


Fig. 1. Graph structures which can be annotated and visualized by means of VAR.

2 Related Work

In addition to the projects already mentioned, annotation in a three-dimensional virtual environment is the subject of various disciplines and research, in which only a few are actually used. A simple annotation tool in virtual environments such as [5,6,8,10] allows the annotation of objects with texts or voice recordings. This may not sound like much at first glance, but it is an essential point in virtual annotation tools. Since entering text without a keyboard is a challenge in itself, there are already innovative solutions available [7]. At the same time there are also annotation systems in the medical [19] and educational field [18]. This work focuses on the aspect of the interface and the implicit annotation of multimodal objects. As far as known, there is still no solution for three-dimensional virtual environments in this context. With VAR that gap will be closed.

3 Technology

The annotation of multimodal (e.g. textual and pictorial data) data is associated with various requirements. In order to meet these requirements, VAR utilizes a graph-based annotation model that covers a wide range of types of information objects as enumerated in Fig. 2. The data structure used by VAR to let users generate examples of this sort is UIMA [9], a data format commonly used in NLP. UIMA enables the creation of flexible annotation schemes, while numerous automatic annotators exist for pre-processing text resources that generate UIMA-compliant documents (e.g. Hemati et al. [11]).

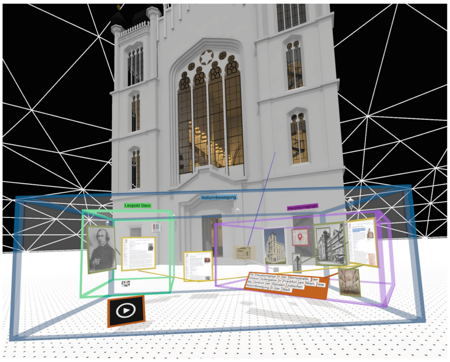
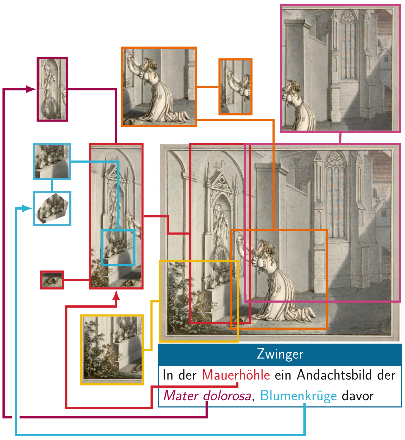


Fig. 2. An overview of a section of multimodal information units (video, audio, geo coordinates, images, 3D models, URLs (visualized as virtual browser windows)), their relations (lines) and their hierarchical encapsulation (large boxes) with VAR. Please note that all 3D objects can also be modeled and entered as independent rooms as well as set in relation to discourse referents in the same virtual environment. In the background the 3D model of the former main synagogue in Frankfurt, which was destroyed during the Nazi regime, is shown. The model was created during a student practical training course [14]. This data is taken from the Stolperwege project [15].



Zwinger

In der Mauerhöhle ein Andachtsbild der Mater dolorosa, Blumenkrüge davor

Fig. 3. Illustration [12] of a scene from Goethe's Faust. Based on the stage instructions on Faust I - Vers 3587 ff., ("In der Mauerhöhle ein Andachtsbild der Mater dolorosa, Blumenkrüge davor" Translation: *In the wall cave, a devotional picture of Mater dolorosa, flower jugs in front of it*), the picture was illustrated according to the artist's imagination. The lines show the segmentation of individual image sections for a more detailed description. The arrows connect the individual text passages with the respective described contents. In this example, individual segments were created recursively from the main image (the example is kept simple for clarity).

To enable the database-oriented processing of such documents, VAR utilizes the *UIMA Database Interface* [2]. In order to enable the annotation of multimodal networks, the following properties are required for VAR's interface:

(a) **Flexibility:** Some applications have to annotate strictly according to a scheme, and others are very flexible. In both cases, however, it is essential that both situations are addressed by the interface. For this purpose, VAR visualizes all relationships and objects based on the classes available in the underlying data model. This allows to limit or extend the annotation capabilities by modifying the data model.

(b) **Intuitive Handling:** In order to enable the usability of an interface, its usage must be simple. However, the use of 3D glasses and the corresponding controllers or other input devices makes this requirement considerably more difficult. To avoid this the VAR combines the possibilities of selecting elements by eye contact, touches by the corresponding controllers or by data gloves. To reduce the learning rate the VAR uses interaction methods borrowed from real life.

(c) **Clarity of presentation:** A brief view on complex graph structures shows that their visualization usually does not contribute to clarity. In order to keep the clarity, the annotations in the VAR are displayed in different ways. This means on the one hand that nodes can be grouped, expanded and hidden and on the other hand that edges can be equipped with different detail levels. In addition, in three dimensional environments this is the most complex component.

(d) **Simple data entry:** Besides the presentation of annotations, the input of data is equally important. In this case VAR basically differs between two options:
   - Multimodal selection of content in VR using virtual browsers and RESOURCES2CITY [13]. The first enables the selection of various contents (text, videos, audio, images) from a virtual browser into the virtual environment (Fig. 2). The second allows the selection of resources on the local system by visualizing the folder structure as a traversable city with buildings for the files.
   - The text input is done via the virtual keyboard which also supports speech to text.

VAR is completely controlled by the virtual hands, which can be used via VR controller or data gloves. These allow the user to interact with the environment and create various objects at any location. As visualized in Fig. 2, a collection of different multimodal objects were created in the virtual environment. Besides the possibility to load the different data types into the virtual environment, they can also be modified, linked or used for further actions. The latter means in VAR that texts and images can be segmented (cf. Fig. 4) and text or image elements from browsers can be selected. Furthermore, all objects can be linked (lines) and grouped (boxes). In order to link objects with each other, a line must be drawn between objects using the *virtual finger*, which can subsequently be attributed. In addition, objects can be grouped hierarchically by

Table 1. Extract of the spatial relationships within an image.

Orientation
in
partOf
inFrontOf
behindOf
aboveOf
belowOf
rightOf
leftOf

creating nodes in which different amounts of objects can be assigned (Fig. 2). At the same time, the editability of the group can be switched on and off to prevent the objects in the grouping from being accidentally ungrouped. The annotations described previously are explicit, in the sense that an object is actively related to another object. There are also implicit annotations, which annotate objects based on spatial relationships. These relationships are established by the annotation process itself and create a meta-annotation, which in this case can only be effective through the three-dimensional annotation environment, since in this case depth perception can be used. As exemplified in Fig. 5, the spatial positions of the previously segmented images are related to each other. Here, the relationship "Orientation:*inFronOf*" is implicitly annotated by the visual positioning of a segment in front of another segment. It is performed by comparing the relative positions of the individual segments on the original image with each other to determine the spatial relationship. In the case of incorrect classifications, these annotations can be adjusted through movement. In addition, all spatial relationships, taken from [4], (see Table 1), which can be annotated more than once (e.g. Object **A** is *inFrontOf* and *aboveOf* **B**). Furthermore, the distinction whether an object is a *in* another or is placed in *frontOf* it is determined by perspective implicit annotation (cf. Figs. 4, 5). Notice that this functionality is currently only implemented in the VR version of VAR. Moreover, this option can be disabled so that an annotator does not accidentally change the implicit annotation of elements when moving them. By default, however, this is enabled.

4 Annotation Scenario

To illustrate the advantages of a 3D tool as exemplified by VAR, we consider a scenario of multimodal annotation taken from the "Faust" [1] project and start with annotating in a 2D fashion. Figure 3 illustrates several content elements of an image and an accompanying text that are linked with each other to manifest, for example, part-whole and intermedial relations. In [3] we described several annotation layers that need to be related in such an annotation scenario. Figure 3 exemplifies an outline of them by using a graph layout in which information objects are linked as vertices by means of colored arcs. Obviously, network-like representations, as shown in this example, quickly become unclear, so that object connections become hardly distinguishable. Any additional annotation (regarding, for example, dependency relations among tokens in the embedded text, positional relations (e.g. *inFrontOf*) of

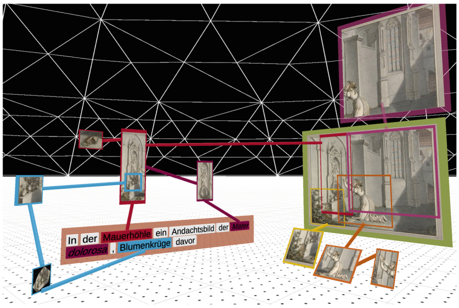

In der Mauerhöhle ein Andachtsbild der Mater
dolorosa , Blumenkrüge davor

Fig. 4. Similar task as annotated in Fig. 3 in the virtual environment of the VAR.

pictorial objects or ontological relations (`Mater dolorosa` *is a* "devotional picture") would render this representation even more unclear. In order to circumvent this problem of *decreasing representational clarity as a result of increasing annotational depth*, VAR uses a 3D representation format that can be manipulated in a multimodal manner: explicitly and implicitly. As illustrated in Fig. 4, VAR generates the same relations as the so called *OWLnotator* [3]. The difference to *OWLnotator* and comparable tools is in the used data model as well as in the 3D visualization and placement of annotations. As mentioned, VAR enables implicit perspective annotations through movements and actions performed by its users. Figure 5 shows this by considering the same annotation scenario as Fig. 3.

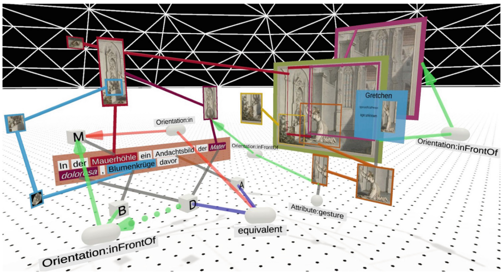


Fig. 5. Extended annotation of Fig. 4. The segmented images are arranged in perspective and their content is related to each other (e.g. pink segment). The segments are related with discourse referents (blue box "Gretchen" as literary person) or get any attributes defined by the existing data model. There are relations between discourse referents (all objects in the VR implicitly become discourse referents) which can be related together (purple line) as a hyper edge to another object (e.g. red line). Implicit relations (green dotted line) can be visualized as well as groupings (DR D). (Color figure online)

The information units in the text to be annotated are represented as discourse referents (cubes) [16,20]. Discourse referents (DR) are conceptual representations of objects as subjects of statements, attributions etc. within a semiotic aggregate (i.e. a text or an image). Starting from a text, visual depictions of DRs (cubes) are connected with the corresponding tokens manifesting them. In the scenario depicted by Fig. 5, `Mauerhöhle` is depicted by the DR M, `Blumenkrüge` by B and the `Andachtsbild` by A. `Mater dolorosa` is a multi-word token that is represented by the DR D. The annotations of Fig. 5 have been created by means of simple controller-based gestures, as explained in Section Technology. It should be noted that the annotation of object relations is partly done by the implicit positioning of virtual objects within the VAR's 3D environment. The relationship `Blumenkrüge` *inFrontOf* `Mater Dolorosa` was annotated, for example, by placing the corresponding representations within the VR. In this way, we take profit from a *fourth dimension* in which perspective topological arrangements of annotation objects are explored to annotated them. From this perspective, it is a relatively small step towards motion-controlled annotations as by-products of the annotator's movements in space. Figure 5 additionally demonstrates the visualization of a subgraph of Type 4 (red line in conjunction with red arrow) (for these types see Fig. 1) and of an implicit, inferential annotation (green dotted arrow): since the `Blumenkrüge` are in *inFrontOf* the

Mauerhöhle and the Mater dolorosa is *in* the Mauerhöhle, the Blumenkrüge are also *inFrontOf* of the Mater dolorosa. Though annotations can be positioned at any desirable location in VR, the format underlying VAR is limited with respect to the amount of information that can be displayed and processed in a graph-like manner. However, moving within such a space allows annotators for freely changing their perspective with respect to focal annotation objects and to use a wider bandwidth of multimodal information processing to manipulate them.

5 Conclusion

In this paper we presented VAR as a tool for graph-based annotations of multimodal objects. Through the use of 3D glasses, this tool enables an immersion of common and complex annotation processes. For this, a scenario for multimodal annotation of image segments was presented in comparison to the 2-dimensional annotation tool *OWLnotator* in which the three-dimensional annotation environment provides considerable advantages. These advantages become evident when not only the representation of multimodal information units is explored, but also the possibility of implicit annotation through the movement of the annotators. Although the graphical representation of complex correlations can still be extensible, a considerable complexity can be represented with the previous implementation. For this reason, the next development steps will include the implementation of a graph algorithm to automatically position the elements of a graph and the automatic analysis of images. It is also important to identify the scope to which the principle of implicit annotation of relations among objects can be used and implemented in augmented reality. Anyway, the development and exploration of annotation and visualization methods in three-dimensional virtual environments has just begun.

References

1. Abrami, G., Freiberg, M., Warner, P.: Managing and annotating historical multimodal corpora with the eHumanities desktop - an outline of the current state of the LOEWE project Illustrations of Goethe's Faust. In: Historical Corpora, pp. 353–363 (2015)
2. Abrami, G., Mehler, A.: A UIMA database interface for managing NLP-related text annotations. In: 2018 Proceedings of LREC, LREC 2018, 7–12 May, Miyazaki, Japan (2018)
3. Abrami, G., Mehler, A., Pravida, D.: Fusing text and image data with the help of the OWLnotator. In: Yamamoto, S. (ed.) HIMI 2015. LNCS, vol. 9172, pp. 261–272. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20612-7_25
4. Bateman, J.A., et al.: A linguistic ontology of space for natural language processing. Artif. Intell. **174**(14), 1027–1071 (2010)
5. Bowman, D.A., Hodges, L.F., Bolter, J.: The virtual venue: user-computer interaction in information-rich virtual environments. Presence **7**(5), 478–493 (1998)

6. Brown, R.A.: Conceptual modelling in 3D virtual worlds for process communication. In: 2010 Proceedings of APCCM, APCCM 2010, Brisbane, Australia, pp. 25–32. Australian Computer Society Inc. (2010)
7. Dudley, J.D., Vertanen, K., Kristensson, P.O.: Fast and precise touch-based text entry for head-mounted augmented reality with variable occlusion. ACM Trans. Comput.-Hum. Interact. **25**(6), 30:1–30:40 (2018)
8. Gabbard, J.L.: A taxonomy of usability characteristics in virtual environments. Ph.D. thesis, Virginia Tech (1997)
9. Götz, T., Suhre, O.: Design and implementation of the UIMA Common Analysis System. IBM Syst. J. **43**(3), 476–489 (2004)
10. Harmon, R., et al.: The virtual annotation system. In: 1996 Proceedings of the IEEE Virtual Reality Annual International Symposium, pp. 239–245. IEEE (1996)
11. Hemati, W., Uslu, T., Mehler, A.: TextImager: a distributed UIMA-based system for NLP. In: Proceedings of the COLING 2016 System Demonstrations. Federated Conference on Computer Science and Information Systems, Osaka, Japan (2016)
12. Frankfurter Goethe-Haus/Freies Deutsches Hochstift. Gretchen vor der Mater dolorosa (2015). https://hessen.museum-digital.de/index.php?t=objekt&oges=2058&navlang=de
13. Kett , A., et al.: Resources2City Explorer: a system for generating interactive walkable virtual cities out of file systems. In: Proceedings of the UIST 2018, Berlin, Germany (2018)
14. Kühn, V., et al.: Digital reconstruction of the former main synagogue in Frankfurt. Result of a student practical course Ubiquitious Texttechnologies in summer term 2018 (2018)
15. Mehler, A., et al.: Stolperwege: an app for a digital public history of the holocaust. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT 2017, Prague, Czech Republic, pp. 319–320. ACM (2017)
16. Mehler, A., et al.: VAnnotatoR: a framework for generating multimodal hypertexts. In: Proceedings of HT 2018, Baltimore, Maryland. ACM (2018)
17. Mehler, A., et al.: Wikidition: automatic lexiconization and linkication of text corpora. Inf. Technol. **58**, 70–79 (2016)
18. Renner, P., Pfeiffer, T.: Evaluation of attention guiding techniques for augmented reality-based assistance in picking and assembly tasks. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion, IUI 2017 Companion, Limassol, Cyprus, pp. 89–92. ACM (2017)
19. Saalfeld, P., Glaßer, S., Preim, B.: 3D user interfaces for interactive annotation of vascular structures. In: Mensch und Computer 2015-Proceedings (2015)
20. Spiekermann, C., Abrami, G., Mehler, A.: VAnnotatoR: a gesture-driven annotation framework for linguistic and multimodal annotation. In: Proceedings of the AREA Workshop, AREA, Miyazaki, Japan (2018)
21. Teo, T., et al.: Data fragment: virtual reality for viewing and querying large image sets. In: 2017 IEEE Virtual Reality, January, pp. 327–328 (2017)
22. Zhao, J., et al.: Annotation graphs: a graph-based visualization for meta-analysis of data based on user-authored annotations. IEEE Trans. Vis. Comput. Graph. **23**(1), 261–270 (2017)