A Study of Narrative Creation by Means of Crowds and Niches

Oana Inel
Vrije Universiteit Amsterdam
oana.inel@vu.nl

Sabrina Sauer
University of Groningen
s.c.sauer@rug.nl

Lora Aroyo
Vrije Universiteit Amsterdam
lora.aroyo@vu.nl

Abstract

Online video constitutes the largest, continuously growing portion of the Web content. Web users drive this growth by massively sharing their personal stories on social media platforms as compilations of their daily visual memories, or with animated GIFs and memes based on existing video material. Therefore, it is crucial to gain understanding of the semantics of video stories, *i.e.*, what do they capture and how. The remix of visual content is also a powerful way of understanding the implicit aspects of storytelling, as well as the essential parts of audio-visual (AV) material. In this paper we take a digital hermeneutics approach to understand what are the visual attributes and semantics that drive the creation of narratives. We present insights from a nichesourcing study in which humanities scholars remix keyframes and video fragments into micro-narratives *i.e.*, (sequences of) GIFs. To support the narrative creation for humanities scholars a specific video annotation is needed, *e.g.*, *(1)* annotations that consider literal and abstract connotations of video material, and *(2)* annotations that are coarse-grained, *i.e.*, focusing on keyframes and video fragments as opposed to full length videos. The main findings of the study are used to facilitate the automatic creation of narratives in the digital humanities exploratory search tool DIVE+[1].

Introduction

Social media provide a mainstream environment to produce, share and comment on video material, which constitutes the largest and still growing portion of Web content (CISCO 2016). An increasingly popular form of shared content are GIFS (Bakhshi et al. 2016) as *micro-stories*, *i.e.*, short video fragments that contain summaries or highlights of video content on participatory platforms GIPHY and Twitter Vine or social media platforms such as Facebook and Instagram.

Humanities scholars use AV archives (De Jong, Ordelman, and Scagliola 2011) to answer their research questions (Melgar et al. 2017), but they face the challenge of grappling with a vast amount of diverse AV content. The DIVE+ (De Boer et al. 2015) tool is conceived to assist scholars in their exploration of digital content to ultimately create meaningful stories and narratives. DIVE+ extends the digital hermeneutics approach (Van Den Akker et al. 2011) by providing interactive access to multimedia objects enriched with events, people, locations and concepts.

Visualizing, mapping and constructing narratives play a significant role in humanities research as they help to contextualize historical material (de Leeuw 2012; Mamber 2012). The remix of AV content as animated GIFs (Highfield and Leaver 2016) gained popularity as an object of study and it is considered a powerful way of understanding the implicit aspects of storytelling. However, the availability of metadata information and semantic annotations (Maccatrozzo et al. 2013; Aroyo, Nixon, and Miller 2011) such as events, objects depicted in the video, relevance of the videos is still a fundamental requirement (Kemman et al. 2013) for scholars to accelerate their narrative-formation process.

The focus of this paper is to understand how niches (De Boer et al. 2012), humanities scholars, interact with AV archives to generate (micro-)narratives. Our research question is: *can we model the data and the semantics of AV content to ease the creation of narratives?* To answer this question we conduct a nichesourcing study with millenials, humanities students in which they use AV content to create stories by means of sequences of GIFs. We analyze the narrative creation process on three levels: *(1)* data - the remixed videos to understand how the story is developed, *(2)* narrative - the micro-story created in and across sequences of GIFs to understand what drives the creation of a narrative, and *(3)* semantics - the keywords describing the story to understand the data enrichment needed to generate narratives.

On the Use of Narratives in Digital Humanities

DIVE+ accommodates the digital hermeneutics approach by means of proto-narratives, *i.e.*, relations between events and their participating entities. To support the creation of such proto-narratives, we gathered events and links between their participating entities in textual AV content (*i.e.*, description) through a hybrid machine-crowd pipeline (de Boer et al. 2017). To further improve the narrative exploration and creation in DIVE+, we performed a nichesourcing study with millennial digital humanities master students to understand how this community builds stories using AV material and which are the needs in terms of data representation. While in previous studies we focused on textual AV content, the current study aims to understand the creation of narratives through visual aspects such as video stills and fragments.

Copyright © 2018 for this paper by its authors. Copying permitted for private and academic purposes.

[1] http://diveplus.beeldengeluid.nl/

Nine international humanities master students (age between 21-25) enrolled in an interdisciplinary course about urban street visualization in Amsterdam participated in our niche study. Their **task** was to explore a dataset of archival AV material and to construct overarching micro-stories, in the shape of sequences of GIFs. A GIF is composed of three keyframes, or a (set of) short video fragment(s). The students were free to explore the dataset and to create GIFs about topics that drew their attention in relation to the city of Amsterdam, or in relation to the course literature.

The **dataset** consists of archival video material about Amsterdam, part of the Netherlands Institute for Sound and Vision[2] (NISV) open collections. We retrieved 624 videos created between 1910-1989 on the NISV portal using the search keyword "Amsterdam". The dataset consists of news broadcasts, varying in length from 50 seconds to 10 minutes, from which we identified three time periods, as shown in Table 1.

Table 1: Dataset overview

Time Period	Period Interval	#Videos	#Users
P1	1910-1929	60	2
P2	1950-1969	288	3
P3	1970-1989	96	4

In the **study** we asked the students to choose a time period in Table 1 and to watch at least 20 videos from that period. The users had one week to complete the entire task, to log their activity[3] and: *(1)* indicate the *GIF type*, *i.e.*, keyframe- or fragment-based; *(2)* describe each GIF, keyframe and video fragment with keywords; *(3)* provide the *timestamps* of the keyframes (keyframe-based GIFs) or the *interval* of the video fragment (fragment-based GIFs), among others. The students were also asked to prepare a short presentation to describe and motivate *(1)* the videos and the time period they selected, *(2)* the selection of keyframes and video fragments and *(3)* the story that is told in their GIFs.

Nichesourcing Study Results

We present the study results[4] and analyze the data gathered from the participating users by focusing on keyframes, video fragments, GIFs and finally, the overarching micro-stories.

The Data Level

The users picked a time period as shown in Table 1. Their choice was informed by either: *(1)* feeling unknowledgeable about that period or *(2)* curiosity about a period when their parents were their own current age. In total, 68 videos were used across all the micro-stories and seven videos were used in more than one micro-story. All the overlaps occurred for the users that chose period P3, which is explained by the low number of videos in P3 and the fact that the users were asked to watch at least 20 videos. On the average, each user used eight videos to generate a story, with a minimum of three and a maximum of 20 videos per story.

Each story was composed of around eight GIFs (stdev of five GIFs), with a minimum of four and a maximum of 20 GIFs. In total, 75 GIFs were generated: seven keyframe-based GIFs and 68 fragment-based GIFs. Only two users generated keyframe-based GIFs, while all nine users generated fragment-based GIFs. The 68 fragment-based GIFs were generated by remixing and combining 89 video fragments, meaning that around 25% of the fragment-based GIFs were composed of more than one video fragment. On average, 10 video fragments (stdev of 10) were used in each micro-story, with a minimum of two and a maximum of 35 video fragments. Furthermore, eight GIFs were generated by remixing keyframes and video fragments from multiple videos (six keyframe-based and two fragment-based GIFs).

In general, mostly keyframes and fragments from the beginning of the videos were picked (55.45%), followed by keyframes and fragments from the middle (24.55%) and then by keyframes and fragments from the end of the video (20%). When multiple keyframes and fragments from the same video were remixed in the same GIF, the *order was always preserved*, *i.e.*, the keyframes and the fragments were used in chronological order with respect to the video stream. However, when looking at the entire story, we observe that the users *break the natural temporal and linear sequence of videos* by starting the story with video fragments or keyframes from the middle or the end part of the videos.

The majority of the GIFs are shorter than six seconds, with only a few longer than 10 seconds. The average length of a story is 43 seconds, with a maximum length of one minute and 48 seconds and a minimum length of 12 seconds. On average, only 3.6% of the videos length was used to generate each story, but, the length of the story is not always proportional with the total length of the videos.

The Narrative Level

The users focused their micro-stories around themes that were either inspired by the content of the videos, or by the course literature (*i.e.*, visualization of urban spaces). The themes of the stories are: *(1)* mobility across the city, *(2)* citizens co-constructing urban spaces, *(3)* gender relations and *(4)* how urban routines relate to feelings of alienation in a globalized world. Some users created literal narratives, depicting aeroplanes, trains and bicycles to indicate mobility, while others worked on an abstract level by, for example, juxtaposing fragments of a person in a deep-sea diving suit with shots of a newspaper article lamenting loneliness in the city, to create a story about alienation.

Users reported that creating sequences of GIFs enabled them to develop more elaborate stories. However, moving from GIF to GIF *does not denote a sequential development in time*, but it is used to *zoom out* spatially, or to create a jarring contrast between GIFs and thus, a more abstract story - for example, moving from a GIF about riots in the street, to a deserted, ruined square in the city, to children repainting a building, to create a story about urban decay and ideals. Similarly, the story about gender relation creates a counterpoint between women undergoing beauty procedures, while men, in a separate GIF, seemingly loom over them.

[2] http://www.beeldengeluid.nl

[3] Log File Template: http://tinyurl.com/zwgotp7

[4] https://tinyurl.com/alternate-stories

The Semantics Level

The users were asked to provide keywords, tags, for their GIFs, selected keyframes and video fragments. These tags represent the users' interpretation of the multimedia content comprising their narratives and do not necessarily describe the content, but act as an interpretation medium for the story. To determine the type of keywords, we manually evaluated them using the Panofsky-Shatford model (Panofsky 1962; Shatford 1986) presented in (Gligorov et al. 2011). We distinguish three levels of keywords: *abstract* - symbolic or subjective concepts that allow for various interpretations, *general* - generic words and *specific* - property of being unique. Further, each level consists of four facets: *who* - subject, *what* - object or event, *where* - location and *when* - time.

We classified 207 (168 unique) tags that describe the GIFs and 262 (159 unique) tags that describe the keyframes and fragments composing the GIFs. The majority of the keywords are *general*, followed by *specific* and then by *abstract* keywords. When looking at the facets, we observe that more than 60% of the keywords belong to the *what* facet. The smallest number of keywords belongs to the *when* facet, with around 1% in all cases. While the keywords describing the *who* and *where* facets are evenly distributed among the keywords describing the GIFs, the amount of keywords describing the keyframes and fragments belonging to the *where* facet is much greater than the amount of keywords describing the *who* facet. While at the *abstract* and *general* levels a significant amount of keywords belong to the *what* facet, at the *specific* level, the users provided more keywords belonging to the *where* facet, and less for the *what* facet, showing that users tend to provide specific locations.

In storytelling, people can refer to concepts, perspectives, opinions that are not physically present in the video, but are referred to or expressed. As research (Trant 2009; Gligorov et al. 2010) indicates, there is also a gap between professional and lay user tags describing video content. To understand the semantics of the keywords provided by users, we look at their overlap with: *(1)* the machine extracted keywords and *(2)* the professional tags. We retrieved the professional tags from the NISV portal and we extracted the visual tags and concepts from each video fragment and keyframe composing each GIF using the online tool Clarifai[5], which performs both image and video concepts recognition.

The overlap between the visual and the keywords provided by the users is quite low: 33% with the keywords describing stills and fragments and 49% with the keywords describing the GIFs. At the level of *general* concepts, the tags provided by the scholars overlap in proportion of 99% with the visual tags. This suggests that for (micro)narrative creation, what is visualized - generally - steers the narrative contained in the story. The overlap between user and professional keywords is even lower, 26% for keyframes and fragments and 30% for GIFs. In contrast to the visual tags, the professional tags do contain *specific* tags which usually refer to places, the *where* facet. For the facet distribution at the *general* level, the proportion of overlapping *what* facets is higher at the level of sequences but lower at the level of the GIFs when compared to the visual tags. The professionals-user gap is clearly defined at the level of *abstract* concepts.

[5] https://www.clarifai.com

Discussion and Future Work

The nichesourcing study aimed to bring insight into storytelling in digital humanities by exploring the interaction and interpretation of micro-narratives remixed using archival AV content. Overall, users tend to generate GIFs by remixing material positioned in the first part of the videos, disregarding the GIF position in the final produced micro-story. The temporal aspect is even more disrupted when users start their narrative with GIFs that contain keyframes and fragments from the middle and the end part of the videos, or when they finish their story with GIFs containing keyframes and fragments from the beginning of videos. Therefore, the original temporal sequence of the video is not relevant when remixing video footage for creative storytelling.

Users ascribe similar interpretations and meanings to their micro-narratives to those contained in visual tags, while they tag the chosen sequences more in terms of their function as a narrative building-block. Although at the GIF level users ascribe similar meaning to the video material as the professionals, they engage in scholarly interpretation on the keyframe level. Thus, the interpretation of meaning in storytelling is, to some extent, developed serendipitously and as a user- and context-centric development, driven by humanities research interests. Time seems - as our facet analysis emphasizes - less important than the *where* or *what* facets. Hence, people find events and objects the most relevant when building narratives. General keywords referring to events, objects, places and people almost entirely overlap with visual tags. Thus, the understanding of visual aspects, especially event and concept-centric, is needed to steer the story line.

In summary, humanities scholars need rich enrichments of AV datasets to facilitate the creation of narratives. However, storytelling through video remixing is a creative process that can not rely only on visual aspects. Deep semantic enrichment is needed to cover both implicit and explicit video concepts and perspectives. For exploratory-centric tools such as DIVE+ it is crucial to: *(1)* provide easy access to already extracted keyframes and video fragments as opposed to expecting the user to watch full videos; *(2)* provide deep semantic enrichment of keyframes and video fragments focusing on specific and general actors or people, locations, time periods, objects and most importantly events. Events play a central role in narrative development. Since event centrality is already a main aspect of DIVE+, we will focus on also integrating crowd-driven keyframes and video fragments semantics to offer users direct access to relevant information. DIVE+ users should be able to access smaller video granularity of interest and their enrichments, as opposed to watching the entire video and inspecting general video metadata.

Acknowledgements

The research for this paper was made possible by the CLARIAH-CORE (www.clariah.nl) project financed by NWO and by the Netherlands Institute for Sound and Vision and NWO under project nr. CI-14-25.

References

[Aroyo, Nixon, and Miller 2011] Aroyo, L.; Nixon, L.; and Miller, L. 2011. Notube: the television experience enhanced by online social and semantic data. In *Consumer Electronics-Berlin (ICCE-Berlin), 2011 IEEE International Conference on*, 269–273. IEEE.

[Bakhshi et al. 2016] Bakhshi, S.; Shamma, D. A.; Kennedy, L.; Song, Y.; de Juan, P.; and Kaye, J. J. 2016. Fast, cheap, and good: Why animated gifs engage us. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 575–586. New York, NY, USA: ACM.

[CISCO 2016] CISCO. 2016. Cisco visual networking index: Forecast and methodology, 2015 - 2020. `http://tinyurl.com/hd7gd45`.

[De Boer et al. 2012] De Boer, V.; Hildebrand, M.; Aroyo, L.; De Leenheer, P.; Dijkshoorn, C.; Tesfa, B.; and Schreiber, G. 2012. Nichesourcing: harnessing the power of crowds of experts. In *International Conference on Knowledge Engineering and Knowledge Management*, 16–20. Springer.

[De Boer et al. 2015] De Boer, V.; Oomen, J.; Inel, O.; Aroyo, L.; Van Staveren, E.; Helmich, W.; and De Beurs, D. 2015. Dive into the event-based browsing of linked historical media. *Web Semantics: Science, Services and Agents on the World Wide Web* 35:152–158.

[de Boer et al. 2017] de Boer, V.; Melgar, L.; Inel, O.; Ortiz, C. M.; Aroyo, L.; and Oomen, J. 2017. Enriching media collections for event-based exploration. In *Research Conference on Metadata and Semantics Research*, 189–201. Springer.

[De Jong, Ordelman, and Scagliola 2011] De Jong, F.; Ordelman, R.; and Scagliola, S. 2011. Audio-visual collections and the user needs of scholars in the humanities: a case for co-development.

[de Leeuw 2012] de Leeuw, S. 2012. European television history online: history and challenges. *VIEW Journal of European Television History and Culture* 1(1):3–11.

[Gligorov et al. 2010] Gligorov, R.; Baltussen, L. B.; van Ossenbruggen, J.; Aroyo, L.; Brinkerink, M.; Oomen, J.; and van Ees, A. 2010. Towards integration of end-user tags with professional annotations.

[Gligorov et al. 2011] Gligorov, R.; Hildebrand, M.; van Ossenbruggen, J.; Schreiber, G.; and Aroyo, L. 2011. On the role of user-generated metadata in audio visual collections. In *Proceedings of the sixth international conference on Knowledge capture*, 145–152. ACM.

[Highfield and Leaver 2016] Highfield, T., and Leaver, T. 2016. Instagrammatics and digital methods: studying visual social media, from selfies and gifs to memes and emoji. *Communication Research and Practice* 2(1):47–62.

[Kemman et al. 2013] Kemman, M.; Scagliola, S.; de Jong, F.; and Ordelman, R. 2013. Talking with scholars: Developing a research environment for oral history collections. In *International Conference on Theory and Practice of Digital Libraries*, 197–201. Springer.

[Maccatrozzo et al. 2013] Maccatrozzo, V.; Aroyo, L.; Van Hage, W. R.; et al. 2013. Crowdsourced evaluation of semantic patterns for recommendations.

[Mamber 2012] Mamber, S. 2012. Narrative mapping. In Everett, A., and Caldwell, J., eds., *New Media: Theories and Practices of Intertextuality*. Routledge. 145–158.

[Melgar et al. 2017] Melgar, L.; Koolen, M.; Huurdeman, H.; and Blom, J. 2017. A process model of scholarly media annotation. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, 305–308. ACM.

[Panofsky 1962] Panofsky, E. 1962. *Studies in Iconology: Humanist Themes in the Art of the Renaissance*. Harper & Row.

[Shatford 1986] Shatford, S. 1986. Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly* 6(3):39–62.

[Trant 2009] Trant, J. 2009. Steve: The art museum social tagging project: A report on the tag contributor experience. In *Museums and the Web*.

[Van Den Akker et al. 2011] Van Den Akker, C.; Legêne, S.; Van Erp, M.; Aroyo, L.; Segers, R.; van Der Meij, L.; Van Ossenbruggen, J.; Schreiber, G.; Wielinga, B.; Oomen, J.; et al. 2011. Digital hermeneutics: Agora and the online understanding of cultural heritage. In *Proceedings of the 3rd International Web Science Conference*, 10. ACM.