GI_Forum 2022, Issue 2
Page: 66 - 72
Short Paper
Corresponding Author:
flavio.horbach@hs-mainz.de
DOI: 10.1553/giscience2022_02_s66

Methods for Georeferencing Linear and Non-Linear Media Content

Flavio Horbach[1], Dominik Visca[1], Sven Pagel[1] and Pascal Neis[1]
[1]Mainz University of Applied Sciences, Germany

Abstract

Various methods, for example image recognition, speech-to-text algorithms, and natural language processing, can be used to capture location references in linear and non-linear media content. The methods differ in terms of the technologies, procedures or media, such as the audio track or video images. Our investigation, which is based on the metadata of a video, reveals that georeferencing media content is possible, and that, using examples from the ARD Mediathek, the genre and thus the content of a video can influence the results.

Keywords:

georeferencing, media content, image, audio, recognition, data analysis

1 Introduction

Nowadays, media content is no longer distributed and consumed only through linear TV and terrestrial broadcast. In the age of 5G and non-linear media libraries such as those of ARD Mediathek (2022), Amazon Prime (2022) or Netflix (2022), the user-specific distribution of media content plays an even greater role than it did ten years ago. Users are able to search for and consume specific content at will, e.g. by searching for keywords, actors or genres. Spatial references, however, which are ubiquitous in most media content, have not yet attracted any significant attention. Creating the spatial references of media content could bring about a paradigm shift: users would be able to search for content via those spatial references and, most importantly, specific content could be recommended according to their current or future locations.

This paper compares various methods of detecting spatial information in linear and non-linear media content, in the form of videos and their metadata, and subsequently georeferencing them.

2 Methodology

Like regular libraries, media libraries are currently undergoing change. Not only the content itself but also its contextual associations have become part of analyses (Müller & Schmunk,

2019). This allows consumer behaviour to be analysed, resulting in better content recommendations. In the case of videos, the supposed location of the action and the actual location where it was filmed become relevant in order to analyse the media content in a more targeted way. However, this information is rarely part of the metadata, and special algorithms are needed to create this spatial reference (Han et al., 2014; Neis, 2021). In order to implement georeferencing (Hackeloeer et al., 2014), various approaches exist which are based on the properties and characteristics of the physical media (Dürscheid, 2005). For example, a video can be analysed by its visual properties or by its audio properties. The analysis of metadata obtained from video properties such as video resolution, length or frame rate. This may in part remove the necessity of analysing the physical (i.e. visual or audio) properties, since the metadata are already available in a usable form.

Here, we review briefly various individual methods for detecting spatial references or locations in linear and non-linear media content. Generally, the actual geocoding (i.e. converting indirect and textual spatial references into coordinates) must be performed after data-collection. Initially, geocoding was primarily understood as generating coordinates from addresses. Today, however, it usually includes more extensive operations, such as the resolution of city names, rivers or mountain ranges to coordinates (Goldberg et al., 2007). Software or services such as the Google Geocoding API (Google Geocoding API, 2022) or the OpenStreetMap Geocoding API (Nominatim, 2022) are able to resolve names and addresses into coordinates. Subsequent spatial analysis of the georeferenced content is possible only when coordinates are known. A location in a video can be a point, a line or a polygon. With the help of these geometries, all types of places, such as cities, countries, roads and POIs, can be mapped.

2.1 Manual data acquisition by humans

The simplest way to georeference videos is to manually gather the spatial references. Here, both the visual and the audio properties are analysed: the locations seen or mentioned can be recognized directly. This approach, of manually collecting the spatial references, offers a simple solution to what is actually a complex problem, because relying on recognition by humans can lead to discrepancies in the results. Each individual has a different knowledge base, which may result in different outcomes for each iteration. For example, an individual with knowledge of Cologne Cathedral is more likely to be able to identify and therefore collect spatial references to the city of Cologne. In contrast, a person who has never seen Cologne Cathedral before will be unable to derive the same spatial reference based on this image. This can lead to inconsistencies in the analysis and subsequent evaluation. In addition, the analysis time is very dependent on the length of the medium (video or audio) which has to be analysed.

2.2 Automatic analysis of audio properties

Another way to georeference videos is to extract information from the audio source and georeference that. Various methods can be used to convert the audio features into text. Many of these tools, such as Google Speech-to-Text recognition (Google Cloud Speech-to-Text, 2022), IBM Watson AI (IBM Watson Speech to Text, 2022) or Microsoft Azure Speech Services (Microsoft Azure Speech Services, 2022), can be accessed via an API. On the provider's servers, the audio data is analysed and converted into text by Speech-to-Text

algorithms. The text can then be interpreted using a variety of Natural Language Processing (NLP) tools. The process of extracting locations, countries or cities from text uses Named-Entity-Recognition. As interest in extracting locations from text has been growing steadily in the field of Spatial Humanities (Won et al., 2018), there are more and more algorithms that are suitable for this purpose. For example, the Spacy (SpaCy, 2022), GeoText (GeoText, 2022) or Flair (FlairNLP, 2022) libraries can be used to extract individual locations from text.

2.3 Automatic analysis of visual properties

Automatic analysis of visual features makes use of the properties and characteristics of the video source. Since a video is a sequence of images, each image has to be analysed individually. Essentially, spatial references or geoinformation can be derived from these individual images using three types of approach.

Object recognition

Artificial intelligence methods are used for object recognition. The traditional process of object recognition can be broken into three stages (Zhao et al., 2019). In the first stage, the part of an image that contains the information about an object is selected. It makes sense to scale the area of the object detected to the actual image size, in order to maximize the potential outcome. The next stage is to extract individual features, i.e. objects in the image. In the last stage, the objects detected are classified. There are already a large number of pre-trained models for this, which are specialized in different object types (Zhao et al., 2019).

Text recognition

The recognition of written text in images is, along with object recognition, one of the most important areas in the field of computer vision. Like object recognition, the recognition of written text is also a multistage process, which closely resembles the stages of object recognition. First, the area to be analysed must be specified. However, large datasets are necessary to train a model that can recognize the position of any written text in an image and later classify it (Zhang et al., 2018).

Pixel-based analysis

Another method for georeferencing images is the analysis of individual pixels of an image. The project PlaNet, for example, trained a neural network with over 2 million images that were already tagged with spatial information (Weyand et al., 2016). Predictions can be made for other images about their location, based on the similarity of their pixels to those of images in the trained model.

3 Analysis

For a prototypical implementation, videos were selected from the ARD Mediathek, and manual and automatic analyses of them were carried out. The ARD Mediathek contains over 200,000 videos in at least 28 different genres (Guides, Politics, Movie, Series, Society, Knowledge or Comedy, for example). In choosing 20 videos, an effort was made to select one

or two from each of as many different genres as possible. The automatic analysis was carried out by evaluating the subtitles of the videos and by using Flair's 'Automatic analysis of audio properties', which is based on Spacy (FlairNLP, 2022). In the manual analysis, both the image and audio sources were analysed by multiple trained people. Initial results showed that, especially in the genres Education, Culture, Documentary and Crime, at least five different spatial references could be identified per video. The length of the selected test videos was between 5 and 80 minutes.

Figure 1 shows a sample of the locations georeferenced using the manual method (green) and the automated analysis (magenta). Table 1 contains the results of the two analyses with subsequent automated georeferencing. It is clear from both Figure 1 and the last column of Table 1 that a large proportion of the spatial references identified using the manual method (about 54% on average) were also detected using the automatic analysis.

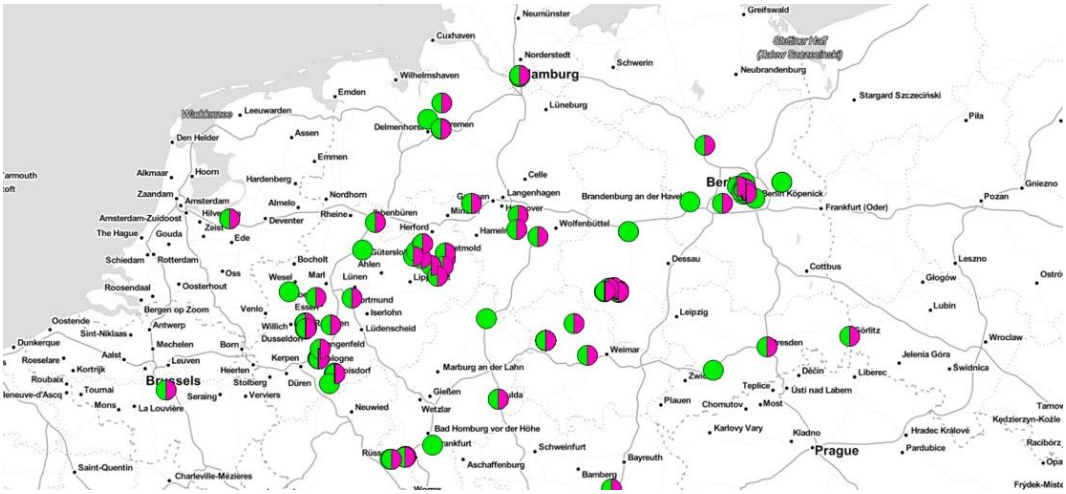

Figure 1: Sample of georeferenced locations from the ARD Mediathek videos, Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under OdbL.

The median number of results found using the automatic procedure was 91% of those found using the manual method. A closer look reveals that this percentage is influenced by cases where the automatic method failed to identify a single spatial reference. However, the opposite can also be seen in the results (i.e. cases where the automatic method identified multiple spatial references). This variable performance will be evaluated in future research using a larger sample size.

Table 1: Number of spatial references identified per test video for each method

No.	Genre / Category	Manual method	Automatic method	Non-identified spatial references	Automatic identified spatial references in manual results
1.	News	4	2	50%	100%
2.	Culture and History	15	9	40%	100%
3.	News	7	1	86%	100%
4.	Documentary	23	17	26%	76%
5.	Documentary	87	83	5%	91%
6.	Entertainment	13	3	77%	66%
7.	Culture and History	23	20	13%	65%
8.	Documentary	19	10	47%	90%
9.	Other	2	3	0%	50%
10.	News	1	0	100%	0%
11.	News	3	1	67%	100%
12.	Documentary	11	20	0%	100%
13.	Other	22	15	32%	86%
14.	News	61	28	54%	92%
15.	Culture and History	32	18	44%	96%
16.	Entertainment	1	5	0%	100%
17.	Documentary	3	2	33%	100%
18.	News	31	0	100%	0%
19.	Entertainment	27	0	100%	0%
20.	Other	15	8	47%	83%

Column 5 in Table 1 shows the number of non-identified spatial references, indicating that a median of 45% of the spatial references could not be identified by the automatic method in comparison to the manual one. At first glance, this value seems high, but it stems from the different underlying data sources. Since only text based on the audio source was used for the automatic approach, the results differ from those of the manual georeferencing because the manual approach also used the video source.

4 Conclusion and Discussion

This article addresses and summarizes multiple methods of georeferencing linear and non-linear media content. The results of the georeferencing are dependent on the method and the underlying data selected, as demonstrated by the analysis presented here. In cases where no spatial reference points could be determined, there is generally little or no data available about the video content. Therefore, for our follow-up analyses, it makes sense to georeference only videos that have at least subtitles and additional metadata, e.g. the synopsis. However, the first analyses also showed that georeferencing works better for some genres than for others, at least for sources taken from the ARD Mediathek. Documentaries and educational content are particularly suitable for georeferencing, since their structure and topics usually contain more spatial reference points. In contrast, the video and audio content of talk shows, concerts or cooking shows is less suitable because their topics tend to offer only a small number of geographical spatial points.

A larger follow-up study will incorporate the use of a more comprehensive and randomized sample set. The design of this future study should consider whether other methods described here should also be incorporated. Based on this prototypical implementation, further research also needs to address the relevance of the identified locations in the context of other use cases, such as geomarketing and product placement.

Acknowledgements

This research was supported by the Südwestrundfunk (SWR) via the project '5G Media2Go'.

References

Amazon Prime (2022). Amazon Prime. Retrieved Jan 22nd, 2022, from https://www.amazon.com/amazonprime

ARD Mediathek (2022). ARD Mediathek · Videos von Das Erste und den Dritten Programmen · ARD Mediathek. Retrieved Jan 22nd, 2022, from https://www.ardmediathek.de/

Dürscheid, C. (2005). Medien, Kommunikationsformen, kommunikative Gattungen. *Linguistik Online*, 14. https://doi.org/10.13092/lo.22.752

FlairNLP (2022). Python. Retrieved Jan 22, 2022, from https://github.com/flairNLP/flair

GeoText (2022). Geotext extracts country and city mentions from text. Retrieved Jan 22nd, 2022, from https://github.com/elyase/geotext

Goldberg, D. W., Wilson, J. P., & Knoblock, C. A. (2007). From Text to Geographic Coordinates: The Current State of Geocoding. *URISA Journal*, 19(1), 33-46.

Google Cloud Speech-to-Text (2022). Retrieved Jan 22nd, 2022, from https://cloud.google.com/speech-to-text#section-3

Google Geocoding API (2022). Google. Retrieved Jan 22nd, 2022, from https://develpers.google.com/maps/documentation/geocoding/start

Hackeloeer, A., Klasing, K., Krisp, J. M., & Meng, L. (2014). Georeferencing: A review of methods and applications. *Annals of GIS*, 20(1), 61–69. https://doi.org/10.1080/19475683.2013.868826

Han, B., Cook, P., & Baldwin, T. (2014). Text-Based Twitter User Geolocation Prediction. *Journal of Artificial Intelligence Research*, 50. https://doi.org/10.1613/jair.4200

IBM Watson Speech to Text (2021). Watson Speech to Text. Retrieved Jan 22nd, 2022, https://www.ibm.com/de-de/cloud/watson-speech-to-text

Microsoft Azure Speech Services (2022). Retrieved Jan 22nd, 2022, from https://azure.microsoft.com/en-us/services/cognitive-services/speech-services/

Müller, F., & Schmunk, S. (2019). Bedeutung und Potenzial von Geoinformationen und deren Anwendungen im Kontext von Bibliotheken und digitalen Sammlungen. Bibliothek Forschung und Praxis, 43(1), 21–34. https://doi.org/10.1515/bfp-2018-0049

Neis, P. (2021): Informationsvisualisierung von Pressemitteilungen auf Basis von Open Source und Open Data – Am Beispiel von Pressemeldungen der Polizei Mainz. Proceedings 21. Internationale Geodätische Woche Obergurgl 2021.

Netflix (2022). Netflix Deutschland – Serien online ansehen, Filme online ansehen. Retrieved Jan 22nd, 2022, from https://www.netflix.com/de/

Nominatim (2022). OpenStreetMap. Retrieved Jan 22, 2022, from https://nominatim.org

SpaCy (2022). Python explosion. Retrieved Jan 22, 2022, from https://spacy.io

Weyand, T., Kostrikov, I., & Philbin, J. (2016). PlaNet—Photo Geolocation with Convolutional Neural Networks. In *Computer Vision – ECCV 2016*, 9912, 37–55). Springer International Publishing. https://doi.org/10.1007/978-3-319-46484-8_3

Won, M., Murrieta-Flores, P., & Martins, B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, 5, 2. https://doi.org/10.3389/fdigh.2018.00002

Zhang, P., Shi, Z., & Gao, H. (2018). Research on Text Location and Recognition in Natural Images with Deep Learning. Proceedings of the 2nd International Conference on Advances in Artificial Intelligence October 2018, 1–6. https://doi.org/10.1145/3292448.3292452

Zhao, Z.-Q., Zheng, P., Xu, S.-T., & Wu, X. (2019). Object Detection With Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, 99, 1-21. https://doi.org/10.1109/tnnls.2018.2876865