future internet

MDPI

Article

Advanced Techniques for Geospatial Referencing in Online Media Repositories

Dominik Warch, Patrick Stellbauer and Pascal Neis *

Department of Applied Informatics and Geodesy, School of Technology, Mainz University of Applied Sciences, 55128 Mainz, Germany; dominik.warch@hs-mainz.de (D.W.); patrick.stellbauer@hs-mainz.de (P.S.)
* Correspondence: pascal.neis@hs-mainz.de

Abstract: In the digital transformation era, video media libraries' untapped potential is immense, restricted primarily by their non-machine-readable nature and basic search functionalities limited to standard metadata. This study presents a novel multimodal methodology that utilizes advances in artificial intelligence, including neural networks, computer vision, and natural language processing, to extract and geocode geospatial references from videos. Leveraging the geospatial information from videos enables semantic searches, enhances search relevance, and allows for targeted advertising, particularly on mobile platforms. The methodology involves a comprehensive process, including data acquisition from ARD Mediathek, image and text analysis using advanced machine learning models, and audio and subtitle processing with state-of-the-art linguistic models. Despite challenges like model interpretability and the complexity of geospatial data extraction, this study's findings indicate significant potential for advancing the precision of spatial data analysis within video content, promising to enrich media libraries with more navigable, contextually rich content. This advancement has implications for user engagement, targeted services, and broader urban planning and cultural heritage applications.

Keywords: natural language processing; named entity recognition; geocoding; online media repository; geospatial information extraction; image-to-text; audio-to-text

Citation: Warch, D.; Stellbauer, P.; Neis, P. Advanced Techniques for Geospatial Referencing in Online Media Repositories. *Future Internet* **2024**, *16*, 87. https://doi.org/10.3390/fi16030087

Academic Editors: Edson Talamini, Letícia De Oliveira and Filipe Portela

Received: 17 January 2024
Revised: 16 February 2024
Accepted: 29 February 2024
Published: 1 March 2024

1. Introduction

In the digital transformation era, video media libraries have emerged as vast data repositories holding immense potential yet are largely untapped due to their non-machine-readable nature [1]. This limitation poses a significant challenge, particularly for users confined to basic search functionalities reliant on standard metadata such as titles, descriptions, categories, and tags. Video content has witnessed exponential growth, predominantly driven by the ubiquity of the Internet and the advent of streaming services like Netflix and social media platforms like TikTok, YouTube, and Instagram [2,3]. These platforms have popularized video as a medium and highlighted the limitations of current search capabilities in media libraries. The inability to search and locate media based on geographical context impedes content exploration, limiting the scope for user engagement, targeted services, and comprehensive analyses.

The integration of geospatial data into video content unlocks diverse applications, such as semantic searches for location-based media, enhancing the relevance of search results [4]. It also allows for targeted advertising, especially on mobile devices where location contexts are dynamic. Moreover, platforms like TikTok and professional media libraries can tailor video recommendations to users' cultural or regional preferences, enriching the personalization of media consumption [5]. These advancements have broader implications in areas including urban planning and cultural heritage, where insights from geocoded historical content can inform conservation and educational initiatives [6].

Traditionally, georeferencing videos has been an area of interest, but the focus has primarily been on specific sub-areas and use cases, leaving broader, more generalized

approaches somewhat unexplored. The recent surge in artificial intelligence, particularly advancements in neural networks and computer vision, has redefined the landscape of video content georeferencing. However, a modern, multimodal approach that considers video in its entirety—encompassing visual content, audio, subtitles, text within video frames, and metadata—is still missing. It becomes increasingly clear that the quality of geospatial referencing videos is intricately tied to progress in fields such as computer science, artificial intelligence, and natural language processing (NLP). While NLP has reached a level of maturity that significantly augments our capabilities in named entity recognition and extraction, we acknowledge that the current evolution stages of other techniques, like computer vision, also shape our results. This interdependence underscores our role as beneficiaries of advancements in these related disciplines, aiming to adapt their most potent and effective methods to our geospatial tasks. Integrating modern methods from neighboring disciplines into a multimodal approach is critical given that different methods may yield different yet correct geospatial references simultaneously. For example, a video might visually depict one location while the audio narrative references another. Our goal is to harness and integrate different established methods to advance our specific niche of georeferencing video content, a domain which, until now, has yet to leverage the potential of these combined technologies fully.

This gap in research, particularly the absence of a comprehensive, multimodal approach to geospatially referencing videos, presents an opportunity for significant advancement in the field. Recognizing this, our research proposes an innovative combination of established methodologies such as computer vision techniques, natural language processing, and geospatial analysis, aiming to extract geospatial references from videos and associate them with precise geocoordinates. This approach is not just an augmentation of existing techniques; it is a paradigm shift toward a more holistic understanding and utilization of video content in the realm of geoinformatics.

2. Related Work

Traditionally rooted in translating textual descriptions of locations into geographic coordinates, geocoding is a cornerstone in spatial data analysis's edifice [7]. Applying NLP in geocoding has been pivotal, employing techniques such as named entity recognition (NER) to extract and categorize place names from unstructured text data [8,9]. Seminal works in this domain have leveraged NLP to recognize textual location references and transform them into mappable coordinates, forming the basis of text-based geospatial data mining [10–13]. NER's evolution, spurred by expansive datasets, has shifted from rule-based to learning-driven paradigms, embracing the subtleties of human language from regional dialects to location ambiguities. This transformation has been particularly significant for geospatial analysis, allowing systems to transition from simple extraction to the nuanced task of entity linking (EL). EL tasks extend further, tethering extracted entities to corresponding entries in knowledge bases such as Wikipedia, thus enriching the spatial data with semantic depth [14]. Geoparsing extends NLP reach into the spatial realm, extracting geographical coordinates from place names recognized in texts. This dual-step process involves toponym recognition followed by toponym resolution, the latter equating to geocoding [15–18]. The rich suite of tools available for these tasks, such as spaCy and Stanford NLP, cater to various applications, from Geographic Information Retrieval (GIR) to emergency response [19]. The methodologies employed range from rule-based to gazetteer-based, statistical, and hybrid approaches, each with its domain of efficacy [20].

Transitioning from text to video presents unique challenges. Early multimodal georeferencing approaches, such as the bag-of-scenes (BoS) technique and histograms of motion patterns (HMPs), integrated traditional text-based geocoding with visual cues, showcasing initial success [21–23]. However, as technologies advanced, these methods saw reduced adoption, overshadowed by metadata-centric strategies, particularly within social media realms in which user-generated content provides a plethora of metadata ripe for geocod-

ing [24,25]. Based on the same approach, Horbach et al. (2022) [26] also focus on mere metadata analysis but show the basic usability of image and audio analysis.

Since professional media libraries do not have this rich metadata at their disposal, AI-supported methods have come to the foreground in recent research. These advanced approaches have demonstrated their efficacy in video classification and multimodal sentiment analyses utilizing convolutional neural networks (CNNs) [27–32]. Like other areas of artificial intelligence, CNN-based place recognition has made great progress in recent years. Models like OverFeat [33–35], AlexNet [36–41], and VGG16 [42–45] and hybrid approaches [46,47] were utilized with great success. The NetVLAD architecture redefined benchmarks in place recognition when applied to significant datasets like Google Street View [48]. The release of the Google Landmarks Dataset v2 provided a vast benchmark for image retrieval and instance recognition, contributing to the training of robust models for place recognition [49,50].

Furthermore, the advent of Optical Character Recognition (OCR) [51–53] and automatic speech recognition [54,55] has opened new avenues for extracting geospatial data from videos. OCR technology can convert visual information into machine-encoded text, enabling the identification of textual information embedded in video frames, such as street signs or text overlays [56–58]. Similarly, audio-to-text transformation tools can transcribe spoken words within videos, processing them like subtitles to parse location references [59].

3. Methodology

The primary objective of this study is to extract and geocode location references from video files using a combination of different methods. This multimodal methodology involves several stages: acquiring data and extracting location data from the visible image, from text in the image, and from audio and subtitles.

As a reference or comparison data set, the location references of the videos for the automated analysis are first manually extracted from the image, audio, and subtitle sources and provided with time stamps. The hits, misses, and false positives of the automated methods to be subsequently evaluated are determined based on this reference data set.

The individual processing steps, especially the computationally intensive ones, were implemented side by side and are thus scalable. This ensures that our methods are efficient and capable of processing large amounts of data, which is crucial for the potential application of our methodology in real large-scale video media libraries.

3.1. Data Acquisition

Within the scope of this article, we have conducted an analysis of media sources from ARD Mediathek [60]. ARD Mediathek was selected as the primary source for this study due to its status as an open and legally accessible library that allows for the downloading and processing of files. This media library offers a diverse array of content, including documentaries, series, movies, and news, providing a rich dataset for analyzing various formats. The availability of high-quality video streams and associated metadata, such as subtitles, titles, descriptions, and categories, makes it an ideal candidate for comprehensive geospatial analysis. The data acquisition process is designed to interface seamlessly with the ARD Mediathek API for the retrieval of video content. This process efficiently ensures the automatic selection and download of the highest-quality video stream, which is crucial for maintaining the clarity of frames required for precise image-based analysis. In parallel, available subtitles in WebVTT format, along with essential metadata such as titles, descriptions, and keywords, are also retrieved. The system is optimized for batch processing and scalability to enhance efficiency, enabling the simultaneous download of multiple videos and subtitles. This capability is instrumental in compiling a large and varied dataset within a manageable timeframe, especially considering the necessity of the highest available video quality. The data acquisition pipeline, as shown in Figure 1, creates a specific folder structure and naming convention for further analysis. If organized within

the same structure, other sources can similarly be subjected to an automated analysis using the methodology outlined in the subsequent steps.

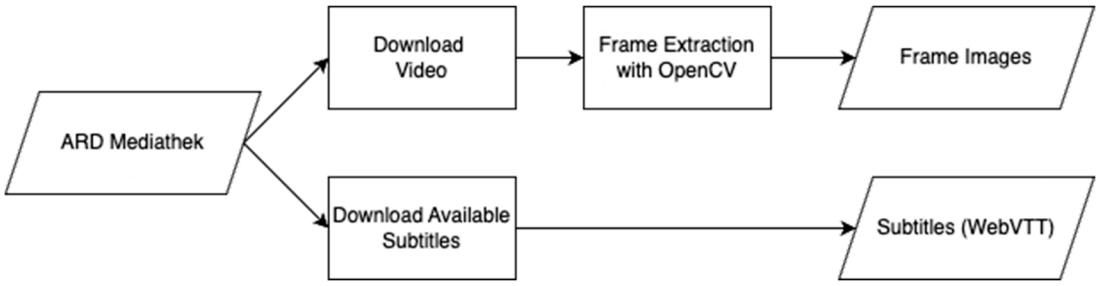


Figure 1. Workflow diagram illustrating data acquisition from ARD Mediathek.

The primary sources of visual information within the video content are twofold: the visible images of the video as a whole and recognizable text within these images, which can appear either as overlayed text (e.g., the names of interview partners, insertions of place names) or as part of the recorded scenes themselves (e.g., recorded street signs). To enable a more precise analysis, the video stream is divided into a sequence of individual images. This discrete splitting of the stream into individual images enables a more detailed and precise analysis of the visual data and, at the same time, makes it more compatible with established image recognition methods.

Following the successful acquisition of videos, the subsequent phase is responsible for extracting images from the downloaded videos, utilizing the image processing capabilities of OpenCV [61]. One key feature of the image extraction process is adaptability in setting the intervals for image extraction, allowing for the extraction of an image each second for the actual automated analysis and every 10 or 30 s for quality control. The extracted images are organized into designated folders corresponding to their source video file and the intended purpose, e.g., analysis or quality control. Like the video and subtitle acquisition, this script is also tuned for large-scale processing to ensure fast but consistent image extraction.

3.2. Analyzing the Visible Image

After the image extraction, we carry out an analysis of the visible image. In this stage, presented in Figure 2, we identify landmarks within the images extracted from the ARD Mediathek videos. Landmark recognition is integral to our methodology as it directly informs the geographical content of videos. By identifying landmarks, we can ascertain specific locations depicted within media, offering a visual anchor point for geospatial analysis. The key player in this process is an advanced machine learning model trained on the Google Landmarks Dataset v2 and further tuned for Europe [49,62], making it highly practical and relevant for our task. The model was selected because it is the only freely available model that was explicitly trained for landmark recognition based on a sufficiently large and recognized corpus.

Before the images are fed into the model, they undergo an additional preprocessing step to prepare them for analysis. This process begins with the initial loading of the image using OpenCV, which typically reads the image in a blue, green, and red (BGR) format. However, since most image processing operations and models, including ours, are optimized for the red, green, and blue (RGB) format, the first step involves converting the image from BGR to RGB. This conversion ensures compatibility with subsequent processing steps and the machine learning model used for landmark recognition.

Once the image is in the correct color space, it is resized to fit the model's input requirements. This resizing is not a straightforward scale-down or scale-up process; instead, it involves a calculated approach to maintain the image's aspect ratio and focal integrity. The script calculates the center of the image and uses this point to determine the dimensions

of a square cropping box, ensuring that the cropped portion retains the most significant parts of the image. After cropping, the image is resized to the target dimensions, set by default to 321 by 321 pixels. This step is crucial as it standardizes the image size across all inputs, a prerequisite for consistent model performance.

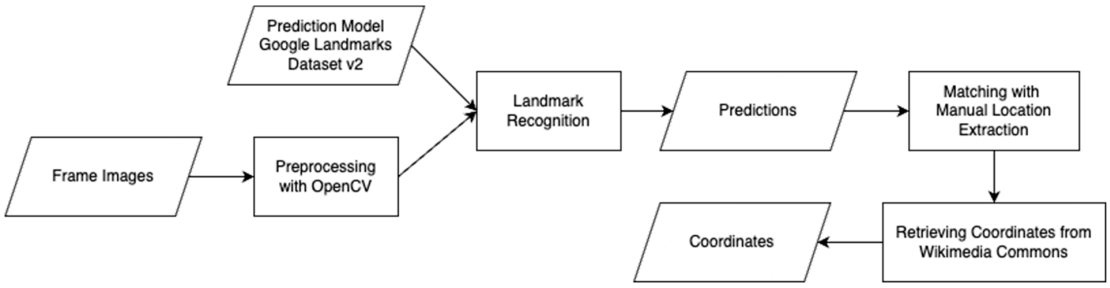


Figure 2. Workflow diagram illustrating the analysis of the visible image.

The final steps in preprocessing involve normalizing the pixel values and preparing the image for the model input. Normalization is conducted by scaling down the pixel intensities to a range between 0 and 1, a common practice in image processing for machine learning to facilitate efficient and stable model training.

As the model processes the images, it generates predictions and confidence scores for each potential landmark. They not only identify the landmarks present in each frame but also provide a measure of confidence in these identifications. This measure is instrumental in assessing the reliability of the recognition. This information becomes even more crucial in our use case as we feed the model an unfiltered selection of all video frames, which may contain only poorly recognizable landmarks. The predictions are stored in a data structure which, in addition to the timeframe, contains a selection of the predictions with the three highest confidence values and the confidence values themselves. This structure allows us to quickly identify the frame images with the same names extracted in the previous phase and precisely enrich further information in subsequent steps. Further geoparsing is unnecessary as the recognized landmarks correspond to specific locations for which coordinates can be retrieved from Wikimedia Commons [63] and added to the referred data structure.

3.3. Analyzing Text in the Visible Image

Extracting text from images is integral to obtaining location data beyond mere landmarks. This dual approach offers a thorough understanding of the location-specific information presented in the media.

We employ the Tesseract OCR engine [51], renowned for its precision in converting diverse image formats into machine-encoded text for text extraction from video frames. Alternatives such as ABBYY FineReader PDF [64] or Adobe Acrobat Pro [65] delivered good results but are challenging to integrate into an automated processing pipeline. Google Cloud Vision API [66], Microsoft Azure Computer Vision API [67], and Amazon Textract [68] allow such integration but are not free to use or are not as flexible and customizable as Tesseract OCR due to their proprietary nature. While Tesseract OCR is robust in handling various image qualities, special attention is given to preprocessing and normalizing the images to mitigate the challenges posed by diverse video formats and resolutions. This normalization includes adjusting contrast and brightness and applying noise-reduction techniques tailored to optimize OCR performance across different video qualities. Once preprocessed, the frames are fed into the Tesseract OCR engine, which scans, detects, and converts text regions into digital text which is subsequently stored in our data structure at the corresponding time frame.

The unstructured text can be analyzed using NLP methods. NER can examine the text for entities such as persons, organizations, and locations and aims to extract the specific names or entities in the text to understand the meaning and structure of the text.

We iteratively perform NER for every frame to process all the contents of the extracted text. The complete process is visualized in Figure 3.

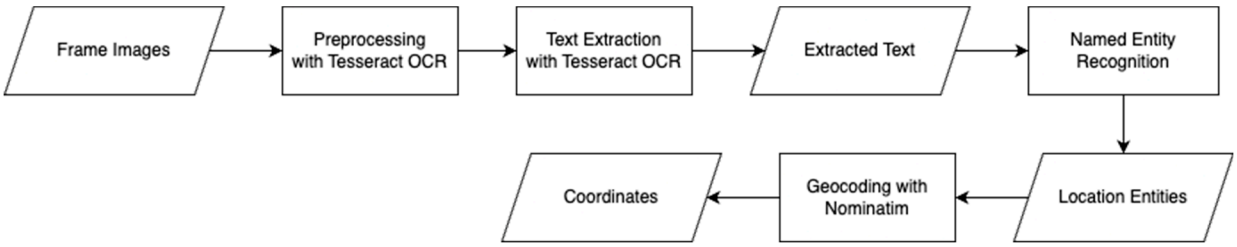


Figure 3. Workflow diagram illustrating the extraction of text from the visible image and performing NER and geocoding to retrieve coordinates.

In the next step, geocoded coordinates are added to these extracted locations. The geocoding of locations plays a crucial role in spatial analysis. A detailed spatial analysis can be performed by assigning geographic coordinates to identify patterns, trends, and relationships between locations. In addition, geocoding enables the visualization of information on maps. The entities of the type Location recognized in the previous step are geocoded by Nominatim [69] based on OpenStreetMap [70] data. If geocoding is successful, a point coordinate pair consisting of latitude and longitude is added to the data structure.

3.4. Analyzing Audio and Subtitles

The audio component of videos is a rich source of data for analysis. These data are converted into text using OpenAI's Whisper, a proven, reliable tool in our transcription process [55]. Whisper accurately captures spoken content, and we additionally embed timestamps to produce transcripts in WebVTT format, which allows for a direct comparison with existing subtitles, ensuring the accuracy and integrity of the transcription and the reuse of our NER method.

We employ the Levenshtein ratio metric, a key component in our evaluation process, to assess the transcription quality by comparing the Whisper-generated entries against the original subtitles [71]. This method enables a differentiated assessment of the accuracy of the transcription against the spoken word. The subtitles are a valuable benchmark in this process due to their alignment with the audio track. After transcription, the text and available subtitles are analyzed using NER and geocoding as previously described. The complete process is shown in Figure 4.

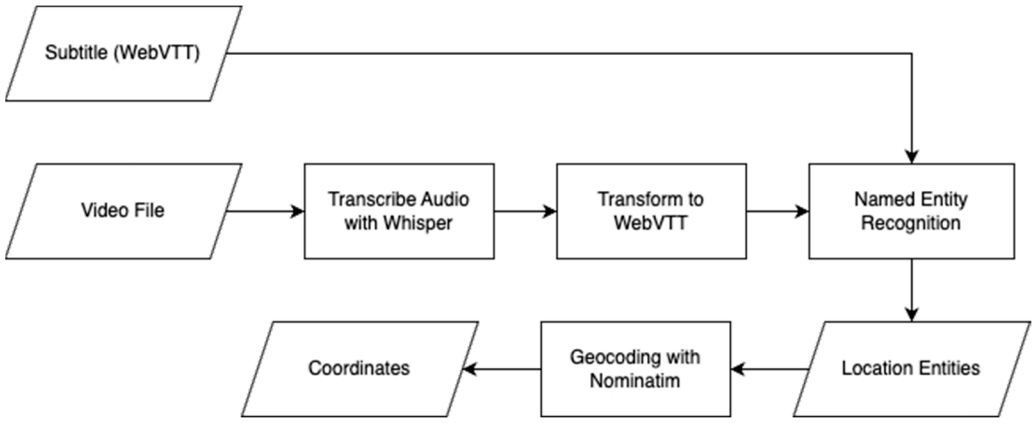


Figure 4. Workflow diagram illustrating the analysis of the audio source and subtitles to retrieve coordinates.

4. Analysis

The methodology described was applied to ten videos from ARD Mediathek with a total of 24,050 extracted frames. Care was taken to evaluate primarily those videos with more frequent location references, i.e., contributions from documentary categories, particularly contemporary city and regional documentaries or history documentaries. Furthermore, only those videos with an available subtitle track were selected for comparison with the transcribed audio tracks. The video length, date of release, and other parameters were randomly selected. We manually determined location references from the videos as a reference value, separated according to the different methods that the automatic information extraction and the NER perform, i.e., the visible image, text in the image, and audio track or subtitles.

An evaluation was conducted to find a suitable tool for NER. A small metadata sample (title and description) of 50 media library entries was compared with different NER tools. Since the focus is on extracting spatial information, the selection of suitable NER tools is based on their performance at extracting location information. Hu et al. (2023) [15] describe a comprehensive comparison of NER tools in their study. Based on these results, NER tools were tested for the toponym extraction of media library entries. The requirements were a simple setup and usability with Python.

- gpt-3.5-turbo: gpt-3.5-turbo is a large language model developed by OpenAI and optimized for chatbots [72]. The prompt was defined to output locations in the text in a standardized and parsable format to extract location references.
- Flair NER: Flair is an NLP framework for facilitating the training and distribution of sequence labeling and text classification. Flair-NER is a standard four-class NER model trained on CoNLL-03. Locations were identified directly using the trained model by querying the LOC tag [73,74].
- GeoTxt: GeoTxt recognizes and extracts location references from text [75].
- spaCy: It is a general NLP tool. The model de_core_news_md was used, and the LOC entities recognized by spaCy were kept as locations [76].
- Stanford Core NLP: This Java implementation of a CRF-based NER was developed and maintained by the Stanford Natural Language Processing Group [77]. It is used via the official Stanza package. The LOC (location) entities recognized by the Stanford NER were retained as locations.
- Stanza: It is a general NLP toolkit with an NER tool built on BiLSTM and CRF, also developed by the Stanford NLP Group [78]. The LOC-type entities were retained as locations. The NER model used is GermEval2014 [79].

As shown in Table 1 the most striking results are that Flair NER has the highest precision and a high F1 score, indicating the accurate recognition of place names. On the other hand, GeoTxt has the lowest precision, the lowest recall, and the lowest F1 score, meaning a poorer overall performance. Stanza shows an excellent overall balance between precision, recall, and F1 score. Based on this evaluation, Flair was used for its performance.

Table 1. Evaluation of named entity recognition tools. Entity used: Location.

Tool	Precision	Recall	F1 Score
gpt-3.5-turbo	0.61	0.79	0.67
Flair NER	0.85	0.80	0.83
GeoText	0.57	0.11	0.19
spaCy	0.63	0.46	0.53
Stanford Core NLP	0.67	0.55	0.60
Stanza	0.78	0.78	0.78

Based on this, the German NER (large model) was used with Flair and trained on CoNLL-03. Locations were identified using this model and could be filtered out from all other recognized entities by querying the LOC tag. The results were each added to a separate column in the previous data structure.

In the 24,050 frames, we manually determined a total of 5262 location references via the audio track or subtitles. NER via Flair has a hit rate of 3262 and 2000 misses. At the same time, we also obtained 676 false positives for which Flair identified location references that our manual classification did not recognize. The results can be rated as good at ca. 62% but can still be improved by minor adjustments. For example, the manual determination of place references is more generous with declined place references or derivational bound morphemes, such as "Österreichs", (Austrian's), "Österreicher", (Austrian; Austrian citizen), and "österreichisch" (Austrian; belonging to Austria) for "Austria". The automated NER tools do not recognize some of those forms as place references, which can be argued for depending on context. However, some misses and false positives are precisely due to these differences. Using a comparison with a Levenshtein distance of 1 instead of an explicit string comparison changes the outcome to 3718 hits—an improvement of 456 or ca. 14%.

When analyzing video frames using OCR technology to identify locations, our results were quite modest. Among the same 24,050 frames, there were just 224 manually identified location references and only 36 had successful identifications. There were 188 instances in which the location should have been identified but was not and 58 instances in which locations were incorrectly identified. This outcome underscores the challenges posed by the dynamic nature of video content, including varying text qualities and complex visual backgrounds.

Utilizing a CNN model for landmark recognition demonstrated a significant trade-off between the breadth of location identification and the accuracy of these identifications. At a high confidence level of 70%, the system identified locations in only 49 instances and missed 3996 location references. Reducing the confidence level to 60% increased correct identifications to 152 but also led to a substantial increase in false positives from 6376 to 31,321. However, the ratio of hits to misses and false positives is not good enough to speak of a good result, regardless of the specific confidence value selected. Initial in-depth analyses of this result showed that differences in designations for one and the same location should not be underestimated. While the previous methods compared location references from Flair, in this case, they were obtained from Flair and labels in the Google Landmark Dataset v2, based on Wikimedia. For example, there are already four different names for the Dresden Academy of Fine Arts: "Kunstakademie", "Lipsius-Bau", "Hochschule für Bildende Künste", and "Kunsthalle Dresden", most of which refer to the same building or building complex.

Across all methods and analysis sources, we manually detected 9531 location references which were detected by automated methods in ca. 40% of cases. The high number of 7110 false positives is mainly due to landmark recognition. The results broken down by method can be seen in Table 2.

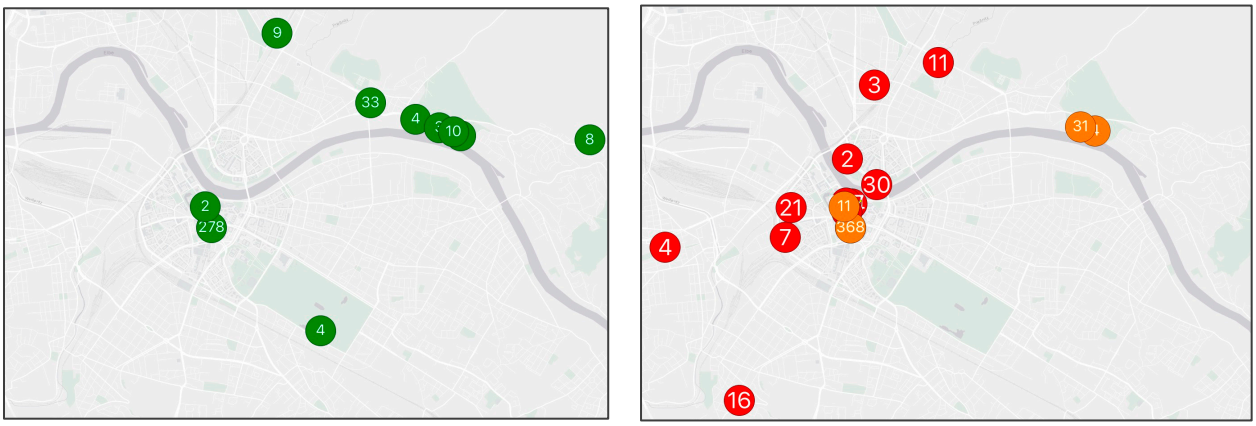
Table 2. Detailed results of different methods for extracting location references.

Method	Precision	Recall	F1 Score
Visible Image (Landmark Recognition)	0.01	0.01	0.01
Text in the Visible Image (OCR)	0.38	0.16	0.23
Audio/Subtitles	0.74	0.71	0.72
Total	0.35	0.40	0.37

Geocoding the 9531 hits and misses and the 7110 false positives worked in ca. 66% of cases, and most failed attempts stemmed from the false positives. On one hand, Nominatim could not geocode all location references with sufficient accuracy or, in some cases, at all;

on the other hand, querying the Wikimedia-based landmark predictions did not return a coordinate in 27% of the cases.

Figure 5 shows an example of location references from the video "Sagenhaft—Dresden". To show the high-resolution accuracy of some of the recognized location references, only one section was selected. Some of the location references (e.g., "China", "Japan", and "New York") are therefore not visible here. On the left side (a), the green circles represent true positives, and the numbers in them are the number of hits for the location. In the city center in particular, there are many hits; this is where the generic "Dresden" is placed by Nominatim. The right side shows false positives in red and false negative (missed) location references.



(**a**) (**b**)

Figure 5. Map excerpt of Dresden showing successfully identified location references in green (**a**) and unsuccessful (orange) location references and false positives (red) (**b**). The numbers represent the number of overlapping location references. Basemap: powered by Esri.

Interestingly, many false positives exist in places that were never recognized in the video in question. The false negatives are concentrated in similar locations to the true positives and suggest that the reliability of the automated procedures is not high enough. The exact location references could sometimes be identified and sometimes not.

5. Discussion

The attempt to analyze visible images within video content using a landmark recognition model has revealed difficulties and limitations inherent in applying AI in geospatial analyses. The model's propensity to identify landmarks with high confidence in images in which none exist is a problem that illustrates the challenges of machine learning interpretability. The black-box nature of AI models often obscures the rationale behind such decisions. As an example, Figure 6 shows a close-up of a person wearing a white hood, which was misidentified as the F 15 Flygmuseum aviation museum in Sweden. This misclassification may be traced to similarities in specific images within the training dataset, which inadvertently biases the model's predictions.

Furthermore, the operational definition of "landmarks" used by the model may be too inclusive, recognizing any image that suggests the presence of a landmark. This broad interpretation can lead to instances in which dominant image features, such as frescoes or crowded events, overshadow the actual landmark. Consequently, the model may fail to recognize the true landmark, hinting at a need for a more nuanced approach to defining and training models on what constitutes a landmark within georeferencing.

(**a**) (**b**)

Figure 6. Misidentification by the landmark recognition model, interpreting a person in a white hood (**a**) as the Swedish F 15 Flygmuseum (**b**), showing challenges with AI interpretability and training data biases. Image sources: (**a**) ARD Mediathek; (**b**) Wikimedia Commons.

Another challenge arises with panoramic images or frames showcasing entire cities or localities. The model used in this study struggles to capture such wide scenes accurately. When multiple landmarks are prominent in the frame, it is likely for the model to recognize one or several; however, when landmarks blend into a cityscape, they often go undetected. This is exemplified by the processed videos featuring city scenes from Dresden, Salzburg, or Hamburg, which the model did not recognize as landmarks due to the absence of these entities as distinct labels in the training data. While the model can identify discrete objects, it falls short at interpreting complex scenes with multiple focal points.

Despite these challenges, the model's ability to accurately recognize a wide array of landmarks, including those that are lesser known or might not be immediately identifiable to an observer with little or no local knowledge, is a testament to its potential. This potential hints at the model's capacity to unveil subtle geospatial details within visual content.

OCR's ability to accurately capture text embedded in video frames, such as street signs or informational plaques, proved inconsistent. Variance in the quality and clarity of these elements within the video often led to unreliable text capture. Factors like motion blur, background complexity, and variable lighting conditions inherent in video content likely contributed to the OCR's performance inconsistencies.

Conversely, text overlays within videos—such as the names of interviewees or credits—were more reliably captured by the OCR technology. These elements typically possess a clearer contrast against the background and a standardized placement, which facilitates more accurate recognition by the OCR algorithms. However, such overlays, as seen in Figure 7a, were not prevalent in our test selection, limiting our ability to assess the OCR technology's efficacy in this domain fully. Future studies could benefit from including a sample of content rich in overlays, like news programs, which frequently employ on-screen text to convey information. Figure 7a shows an instance in which only one ("Festspielhaus Hellerau") of three location references (the others are "Sachsen" and "Dresden") is recognized by OCR. Interestingly, all three location references were not considered during manual acquisition. A title card from which OCR was not able to recognize text can be seen in Figure 7b. In this specific case, a location reference could not be recognized here.

A notable issue was the OCR technology's difficulty with recognizing text presented in large or highly stylized fonts. Title cards, which often showcase artistic typography to capture the viewer's attention, posed a particular challenge.

(**a**) (**b**)

Figure 7. Examples in which OCR captured parts of the text (**a**) and where OCR was not able to recognize text due to large and partly concealed fonts (**b**). Image source: (**a**,**b**): ARD Mediathek.

Audio and subtitle processing has emerged as a promising way forward, highlighting the potential of advanced linguistic models. However, it also presents challenges as the contextual subtleties of location information are highly nuanced. The refinement of transcribed texts using GPT-3.5 Turbo marked a notable step in our methodology, enhancing the meaningfulness and correctness of the data. However, this correction did not notably impact the performance of NER tools. This finding suggests that the effectiveness of NER in recognizing location entities may depend on factors beyond textual accuracy. Given the high fidelity of the audio transcriptions, our approach flexibly treated both audio and subtitles as interchangeable, capitalizing on the extensive textual data they provided. This strategy proved beneficial, particularly as the text volume was well-suited to NER tools designed to process large text corpora.

Regardless, a notable challenge was the contextual interpretation of location references. A reference to a "cathedral" in a video about "Augsburg" implicitly pertains to the "Augsburg Cathedral". With the contextual linkage, the term is more precise for NER tools to handle effectively. Addressing this, one solution could involve integrating additional context through metadata pre-analyses, potentially enriching the locational specificity of such references. However, this integration must be handled carefully as there is a risk of complicating the analysis by distorting or overlaying non-local place references.

6. Conclusions and Future Work

This analysis shows mixed results for enhancing the utility of video media libraries through a multimodal methodology for geocoding video content. It is particularly striking how different the results of the individual methods are and that there are many pitfalls on one hand but also many levers for improving the results on the other. By leveraging advancements in machine learning, natural language processing, and computer vision, we present the foundation of a novel approach to unearthing the rich geospatial data embedded within videos. This elevates the precision of search functionalities and opens new paths for targeted advertising and content personalization.

The exploration of georeferencing within video content reveals a rich seam of potential for future research to advance the precision and applicability of spatial data analysis. Recognizing the varied efficacy of the methods utilized, the primary objective of forthcoming endeavors will be to develop a location scoring system. Such a system would synergize results from various georeferencing approaches to ascertain the relevance of locations depicted or mentioned in videos. By implementing weights in the various methods, the importance and reliability of each georeferencing method can contribute to the most accurate and relevant spatial data points. It is also critical to recognize that the various methods applied—landmark recognition, OCR, and computer vision—may yield distinct yet simultaneously correct geospatial references within the same frame. The location de-

picted visually, the locations mentioned in on-screen text, and the places referenced in dialogue or subtitles can differ. Consequently, an intrinsic error correction that relies solely on the methods described may not necessarily enhance the accuracy of the results. This highlights the need for a nuanced approach to integrating these varied references, which a location-scoring system should aim to address.

In addition, future developments could integrate a more nuanced, context-aware framework for temporal geospatial analyses. This would not only dissect video content into time-coded segments but also understand narrative and thematic progression, offering a dynamic mapping of geospatial references that evolve with the storyline of the video content. Such advancements would not only enhance data granularity but also the interpretability and application of geospatial data in complex video sequences.

The improvement of the landmark recognition model is worth exploring as well. As shown, the current limitations faced due to inaccuracies in the training data of the Google Landmarks Dataset v2 underscore the need for a model that is trained or fine-tuned using a dataset with more precise labeling. This would ensure that landmark predictions are not only accurate but also tailored to the geographical focus of the media being analyzed, thereby circumventing the issue of erroneous predictions that may arise from a more generic dataset.

A surprising result was the quality of the automatic transcription of the audio track to subtitle files and the associated automatic corrections. Although this extends beyond the scope of the present work, further research in this field can contribute to significantly increasing the accessibility of entire media libraries with relatively low personnel and cost expenditures.

Our research has implications that extend beyond the immediate scope of this study. For instance, in emergency response, the real-time georeferencing of video content can provide crucial location data. In environmental monitoring, our methods could assist in identifying and tracking the impact of events across different regions. These potential applications offer a range of opportunities for future research. While a specific roadmap for technological developments is beyond the scope of this article, we are enthusiastic about exploring collaborations with media providers and their media libraries. Furthermore, we recognize the complexity of geospatial data and the challenges it presents. We understand that advances in AI, NLP, and computer vision, each at different stages of development, provide tools of varied efficacy for our geospatial referencing tasks. Our approach seeks to integrate these tools, with an acknowledgment of our reliance on the evolution of these methods to achieve results that meet high expectations. The proposed location-scoring system aims to constructively synthesize these results into a cohesive assessment in future work, thereby enhancing the accuracy and utility of georeferencing video content.

Author Contributions: Conceptualization, D.W., P.S. and P.N.; methodology, D.W. and P.N.; software, D.W. and P.S.; validation, D.W. and P.S.; formal analysis, D.W., P.S. and P.N.; investigation, D.W., P.S. and P.N.; resources, D.W., P.S. and P.N.; data curation, D.W.; writing—original draft preparation, D.W.; writing—review and editing, D.W. and P.N.; visualization, D.W.; supervision, P.N.; project administration, P.N.; funding acquisition, P.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The processed and analyzed videos are publicly available in the ARD Mediathek according to the respective license (https://www.ardmediathek.de). We can provide a list of the specific videos used on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hopfgartner, F.; Schöffmann, K. Interactive Search in Video & Lifelogging Repositories. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, Oslo, Norway, 7–11 March 2017; pp. 421–423.
2. Rupapara, V.; Thipparthy, K.R.; Gunda, N.K.; Narra, M.; Gandhi, S. Improving Video Ranking on Social Video Platforms. In Proceedings of the 2020 7th International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 23–24 July 2020; pp. 1–5.
3. Westphal, C.; Melodia, T.; Zhu, W.; Timmerer, C. Guest Editorial Video Distribution over Future Internet. *IEEE J. Select. Areas Commun.* **2016**, *34*, 2061–2062. [CrossRef]
4. Jamonnak, S.; Zhao, Y.; Curtis, A.; Al-Dohuki, S.; Ye, X.; Kamw, F.; Yang, J. GeoVisuals: A Visual Analytics Approach to Leverage the Potential of Spatial Videos and Associated Geonarratives. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 2115–2135. [CrossRef]
5. Chen, Z.; Shi, C. Analysis of Algorithm Recommendation Mechanism of TikTok. *Int. J. Educ. Humanit.* **2022**, *4*, 12–14. [CrossRef]
6. Hyvönen, E. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web; Synthesis Lectures on Data, Semantics, and Knowledge*; Springer International Publishing: Cham, Switzerland, 2012.
7. Goldberg, D.W.; Wilson, J.P.; Knoblock, C.A. From Text to Geographic Coordinates: The Current State of Geocoding. *Urisa J.* **2007**, *19*, 33.
8. Nadeau, D.; Sekine, S. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investig.* **2007**, *30*, 3–26. [CrossRef]
9. Gelernter, J.; Balaji, S. An Algorithm for Local Geoparsing of Microtext. *Geoinformatica* **2013**, *17*, 635–667. [CrossRef]
10. Leidner, J.L.; Lieberman, M.D. Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *Sigspatial Spec.* **2011**, *3*, 5–11. [CrossRef]
11. Hu, Y.; Mao, H.; McKenzie, G. A Natural Language Processing and Geospatial Clustering Framework for Harvesting Local Place Names from Geotagged Housing Advertisements. *Int. J. Geogr. Inf. Sci.* **2018**, *33*, 714–738. [CrossRef]
12. Stenetorp, P.; Pyysalo, S.; Topic, G.; Ohta, T.; Ananiadou, S.; Tsujii, J. Brat: A Web-Based Tool for NLP-Assisted Text Annotati-on. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 23–27 April 2012.
13. Wang, W.; Stewart, K. Spatiotemporal and Semantic Information Extraction from Web News Reports about Natural Hazards. *Comput. Environ. Urban Syst.* **2015**, *50*, 30–40. [CrossRef]
14. Ling, X.; Singh, S.; Weld, D.S. Design Challenges for Entity Linking. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 315–328. [CrossRef]
15. Hu, X.; Zhou, Z.; Li, H.; Hu, Y.; Gu, F.; Kersten, J.; Fan, H.; Klan, F. Location Reference Recognition from Texts: A Survey and Comparison. *ACM Comput. Surv.* **2023**, *56*, 112. [CrossRef]
16. Gregory, I.; Donaldson, C.; Murrieta-Flores, P.; Rayson, P. Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *Int. J. Humanit. Arts Comput.* **2015**, *9*, 1–14. [CrossRef]
17. Melo, F.; Martins, B. Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Trans. GIS* **2017**, *21*, 3–38. [CrossRef]
18. Leetaru, K.H. Fulltext Geocoding versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia. *D-Lib Mag.* **2012**, *18*. [CrossRef]
19. Gritta, M.; Pilehvar, M.T.; Limsopatham, N.; Collier, N. What's Missing in Geographical Parsing? *Lang. Resour. Eval.* **2018**, *52*, 603–623. [CrossRef] [PubMed]
20. Purves, R.S.; Clough, P.; Jones, C.B.; Hall, M.H.; Murdock, V. Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *FNT Inf. Retr.* **2018**, *12*, 164–318.
21. Li, L.T.; Pedronette, D.C.G.; Almeida, J.; Penatti, O.A.B.; Calumby, R.T.; Da, S.; Torres, R. Multimedia Multimodal Geocoding. In Proceedings of the 20th International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 6–9 November 2012; pp. 474–477.
22. Penatti, O.A.B.; Li, L.T.; Almeida, J.; Da, S.; Torres, R. A Visual Approach for Video Geocoding Using Bag-of-Scenes. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Hong Kong, China, 5–8 June 2012; pp. 1–8.
23. Paris, S.; Halkias, X.; Glotin, H. Beyond SIFT for Image Categorization by Bag-of-Scenes Analysis. In *Pattern Recognition Applications and Methods*; Fred, A., De Marsico, M., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2015; Volume 318, pp. 191–207.
24. Trevisiol, M.; Jégou, H.; Delhumeau, J.; Gravier, G. Retrieving Geo-Location of Videos with a Divide & Conquer Hierarchical Multimodal Approach. In Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, Dallas, TX, USA, 16–20 April 2013; pp. 1–8.
25. Van Laere, O.; Schockaert, S.; Dhoedt, B. Georeferencing Flickr Resources Based on Textual Meta-Data. *Inf. Sci.* **2013**, *238*, 52–74. [CrossRef]
26. Horbach, F.; Visca, D.; Pagel, S.; Neis, P. Methods for Georeferencing Linear and Non-Linear Media Content. *GI_Forum* **2023**, *10*, 66–72. [CrossRef]
27. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* **2013**, 580–587. [CrossRef]
28. Graves, A.; Mohamed, A.; Hinton, G. Speech Recognition with Deep Recurrent Neural Networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.

29. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
30. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Columbus, OH, USA, 23–28 June 2014.
31. Jiang, Y.-G.; Wu, Z.; Tang, J.; Li, Z.; Xue, X.; Chang, S.-F. Modeling Multimodal Clues in a Hybrid Deep Learning Framework for Video Classification. *IEEE Trans. Multimed.* **2018**, *20*, 3137–3147. [CrossRef]
32. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; pp. 439–448.
33. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. OverFeat: Integrated Recognition, Localization and Detec-tion Using Convolutional Networks. *arXiv* **2014**, arXiv:1312.6229.
34. Chen, Z.; Lam, O.; Jacobson, A.; Milford, M. Convolutional Neural Network-Based Place Recognition. *arXiv* **2014**, arXiv:1411.1509.
35. Razavian, A.S.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 512–519.
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
37. Sünderhauf, N.; Shirazi, S.; Dayoub, F.; Upcroft, B.; Milford, M. On the Performance of ConvNet Features for Place Recognition. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 4297–4304.
38. Hou, Y.; Zhang, H.; Zhou, S. Convolutional Neural Network-Based Image Representation for Visual Loop Closure Detection. In Proceedings of the 2015 IEEE International Conference on Information and Automation, Lijiang, China, 8–10 August 2015; pp. 2238–2245.
39. Panphattarasap, P.; Calway, A. Visual Place Recognition Using Landmark Distribution Descriptors. In *Computer Vision–ACCV 2016, Proceedings of the 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016, Revised Selected Papers, Part IV 13*; Springer: Berlin/Heidelberg, Germany, 2016.
40. Bai, D.; Wang, C.; Zhang, B.; Yi, X.; Yang, X. CNN Feature Boosted SeqSLAM for Real-Time Loop Closure Detection. *Chin. J. Electron.* **2018**, *27*, 488–499. [CrossRef]
41. Chen, Z.; Jacobson, A.; Sunderhauf, N.; Upcroft, B.; Liu, L.; Shen, C.; Reid, I.; Milford, M. Deep Learning Features at Scale for Visual Place Recognition. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 3223–3230.
42. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
43. Neubert, P.; Protzel, P. Beyond Holistic Descriptors, Keypoints, and Fixed Patches: Multiscale Superpixel Grids for Place Recog-nition in Changing Environments. *IEEE Robot. Autom. Lett.* **2016**, *1*, 484–491. [CrossRef]
44. Chen, Z.; Maffra, F.; Sa, I.; Chli, M. Only Look Once, Mining Distinctive Landmarks from ConvNet for Visual Place Recogni-tion. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 9–16.
45. Chen, Z.; Liu, L.; Sa, I.; Ge, Z.; Chli, M. Learning Context Flexible Attention Model for Long-Term Visual Place Recognition. *IEEE Robot. Autom. Lett.* **2018**, *3*, 4015–4022. [CrossRef]
46. Radenović, F.; Tolias, G.; Chum, O. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1655–1668. [CrossRef]
47. Kim, H.J.; Dunn, E.; Frahm, J.-M. Learned Contextual Feature Reweighting for Image Geo-Localization. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3251–3260.
48. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN Architecture for Weakly Supervised Place Recogni-tion. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5297–5307.
49. Weyand, T.; Araujo, A.; Cao, B.; Sim, J. Google Landmarks Dataset v2—A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2572–2581.
50. Hausler, S.; Garg, S.; Xu, M.; Milford, M.; Fischer, T. Patch-NetVLAD: Multi-Scale Fusion of Locally-Global Descriptors for Place Recognition. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 14136–14147.
51. Smith, R. An Overview of the Tesseract OCR Engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, PR, Brazil, 23–26 September 2007; pp. 629–633.
52. Islam, N.; Islam, Z.; Noor, N. A Survey on Optical Character Recognition System. *arXiv* **2017**, arXiv:1710.05703.
53. Mittal, R.; Garg, A. Text Extraction Using OCR: A Systematic Review. In Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 15–17 July 2020; pp. 357–362.

54. Malik, M.; Malik, M.K.; Mehmood, K.; Makhdoom, I. Automatic Speech Recognition: A Survey. *Multimed. Tools Appl.* **2021**, *80*, 9411–9457. [CrossRef]
55. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Su-pervision. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022.
56. Sato, T.; Kanade, T.; Hughes, E.K.; Smith, M.A. Video OCR for Digital News Archive. In Proceedings of the 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, Bombay, India, 3 January 1998; pp. 52–60.
57. Saluja, R.; Maheshwari, A.; Ramakrishnan, G.; Chaudhuri, P.; Carman, M. OCR On-the-Go: Robust End-to-End Systems for Reading License Plates & Street Signs. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 154–159.
58. Priambada, S.; Widyantoro, D.H. Levensthein Distance as a Post-Process to Improve the Performance of OCR in Written Road Signs. In Proceedings of the 2017 Second International Conference on Informatics and Computing (ICIC), Jayapura, Indonesia, 1–3 November 2017; pp. 1–6.
59. Paiders, J.; Plume, E. Use of Place Names in the Subtitle Corpus of Highest-Grossing Movies of the Past 20 Years. *J. Int. Symp. Stud. Engl. Croat. Ital. Stud.* **2018**, *1*, 43–60.
60. ARD Mediathek. Available online: https://www.ardmediathek.de/ (accessed on 7 January 2024).
61. Bradski, G. The OpenCV Library. *Dr. Dobb's J. Softw. Tools* **2020**, *25*, 120–125.
62. Cao, B.; Araujo, A.; Sim, J. Unifying Deep Local and Global Features for Image Search. In *Computer Vision–ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part XX 16*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 726–743.
63. Wikimedia Commons. Available online: https://commons.wikimedia.org/wiki/Main_Page (accessed on 7 January 2024).
64. ABBYY FineReader PDF. Available online: https://pdf.abbyy.com/ (accessed on 9 February 2024).
65. Adobe Acrobat: Easily Edit Your Scanned PDF Documents with OCR. Available online: https://www.adobe.com/acrobat/how-to/ocr-software-convert-pdf-to-text.html (accessed on 9 February 2024).
66. Google Cloud Vision API: Detect Text in Images. Available online: https://cloud.google.com/vision/docs/ocr (accessed on 9 February 2024).
67. Microsoft Azure AI Vision Documentation: OCR—Optical Character Recognition. Available online: https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/overview-ocr (accessed on 9 February 2024).
68. Amazon Textract: Automatically Extract Printed Text, Handwriting, Layout Elements and Any Data from Any Document. Available online: https://aws.amazon.com/textract (accessed on 9 February 2024).
69. Nominatim: Open-Source Geocoding with OpenStreetMap Data. Available online: https://nominatim.org/ (accessed on 7 January 2024).
70. OpenStreetMap. Available online: https://openstreetmap.org/ (accessed on 7 January 2024).
71. Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; Vollgraf, R. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In Proceedings of the NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Online, 6–11 June 2019; pp. 54–59.
72. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual String Embeddings for Sequence Labeling. In Proceedings of the COLING 2018, 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1638–1649.
73. Karimzadeh, M.; Pezanowski, S.; MacEachren, A.M.; Wallgrün, J.O. GeoTxt: A Scalable Geoparsing System for Unstructured Text Geolocation. *Trans. GIS* **2019**, *23*, 118–136. [CrossRef]
74. spaCy: Industrial-Strength Natural Language Processing. Available online: https://spacy.io/ (accessed on 7 January 2024).
75. Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 22–27 June 2014; pp. 55–60.
76. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv* **2020**, arXiv:2003.07082.
77. Benikova, D.; Biemann, C.; Reznicek, M. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; pp. 2524–2531.
78. OpenAI GPT-3.5 Turbo Fine-Tuning and API Updates. Available online: https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates (accessed on 7 January 2024).
79. Hossain, M.M.; Labib, M.F.; Rifat, A.S.; Das, A.K.; Mukta, M. Auto-Correction of English to Bengali Transliteration System Using Levenshtein Distance. In Proceedings of the 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, 28–30 June 2019; pp. 1–5.