applied sciences

MDPI

Article

Context-Aware Querying, Geolocalization, and Rephotography of Historical Newspaper Images

Dilawar Ali [1,2,*], Thibault Blyau [1], Nico Van de Weghe [2] and Steven Verstockt [1]

[1] IDLab, Ghent University—IMEC, Technologiepark-Zwijnaarde 122, 9052 Ghent, Belgium
[2] Department of Geography, Ghent University, Krijgslaan 281 (S8), 9000 Ghent, Belgium
* Correspondence: dilawar.ali@ugent.be

Abstract: Newspapers contain a wealth of historical information in the form of articles and illustrations. Libraries and cultural heritage institutions have been digitizing their collections for decades to enable web-based access to and retrieval of information. A number of challenges arise when dealing with digitized collections, such as those of KBR, the Royal Library of Brussels (used in this study), which contain only page-level metadata, making it difficult to extract information from specific contexts. A context-aware search relies heavily on metadata enhancement. Therefore, when using metadata at the page level, it is even more challenging to geolocalize less-known landmarks. To overcome this challenge, we have developed a pipeline for geolocalization and visualization of historical photographs. The first step of this pipeline consists of converting page-level metadata to article-level metadata. In the next step, all articles with building images were classified based on image classification algorithms. Moreover, to correctly geolocalize historical photographs, we propose a hybrid approach that uses both textual metadata and image features. We conclude this research paper by addressing the challenge of visualizing historical content in a way that adds value to humanities research. It is noteworthy that a number of historical urban scenes are visualized using rephotography, which is notoriously challenging to get right. This study serves as an important step towards enriching historical metadata and facilitating cross-collection linkages, geolocalization, and the visualization of historical newspaper images. Furthermore, the proposed methodology is generic and can be used to process untagged photographs from social media, including Flickr and Instagram.

Keywords: article segmentation; cultural heritage data; newspaper; digitization of heritage materials; location-aware computing; historical image rephotography

Citation: Ali, D.; Blyau, T.; Van de Weghe, N.; Verstockt, S. Context-Aware Querying, Geolocalization, and Rephotography of Historical Newspaper Images. *Appl. Sci.* **2022**, *12*, 11063. https://doi.org/10.3390/app122111063

Academic Editor: Valentina Alena Girelli

Received: 26 September 2022
Accepted: 27 October 2022
Published: 1 November 2022

1. Introduction

Cultural heritage institutions have extensive historical collections containing valuable past information. They hold a wide range of items, including photographs, manuscripts, newspapers, historical maps, stamps, and magazines. Newspapers are considered one of the most valuable primary sources for gathering information about a culture, society, or historical period. In the early 1990s, libraries and cultural heritage institutions began to digitize their collections in order to preserve historical materials and provide access to them through web-based interfaces [1–4]. The digitization process generally refers to the annotation of documents and the extraction of historical information using standard optical character recognition (OCR). It is, however, challenging to extract information from newspapers using OCR methods [5–7].

One of the challenges in extracting information from newspapers is dealing with layouts that change over time. Furthermore, dealing with OCR results is also challenging because they are strongly affected by scan quality and the level of noise in historical collections. This makes it more difficult to extract historical data for further analysis [8–11]. An approach such as noise characterization [12] could be used as a preventive measure to limit OCR quality issues and extract historical information more accurately. Aside from the

above-mentioned challenges, the digitized versions of newspapers, including those from KBR, the Royal Library of Belgium, contain only page-level metadata, making it difficult to extract information related to a particular topic. Access to KBR's digitized newspapers is available through the BelgicaPress platform [13]. When someone searches for a term with a keyword, all pages related to that term are displayed. A newspaper page contains numerous articles on different topics. This makes the page-level search less valuable and can lead to irrelevant results. Therefore, it is necessary to improve the metadata to allow a context-aware search in these historical collections.

The next step after digitization and providing access to collections is enrichment. Metadata enrichment is the process of adding extra information about the data. The additional information can be the name of an event, the date of publication, the article number, the author of an article, the picture description, or the location where a picture was taken. The metadata extracted from these newspapers can enhance the collection and thus make it more accessible. In order to explore the collections, traditional image processing techniques are used to segment the articles so that users can perform context-aware searches. Context-aware cultural heritage applications are expected to improve searchability, user interaction, and engagement with historical datasets [14,15].

Computer vision and machine learning techniques are used to further enrich metadata to facilitate article-level search, cross collection linking, named entity recognition, and enable the user to perform context-aware queries on the collections. Based on these advanced computational methods, researchers now have the opportunity to create full-featured virtual interfaces [16–18] for digital humanities research. *NewspAIper* [19], an AI-based metadata enrichment solution for historical newspaper collections, was developed to help humanities researchers extract information from vast collections relevant to specific research scenarios. This tool led to the development of corpora on specific research themes. A corpus could be a collection of articles related to "Strikes in Belgium" or a bundle of pictures related to the "style of buildings".

The next step is the analysis of the data corpus after it has been compiled according to specific research scenarios. Among cultural heritage materials, text analysis [20,21] has been the main focus of research for decades. In the era of increasing data availability, technologies such as linked open data and the semantic web are becoming more popular for querying and exploring textual information [22–24]. The lack of focus on image analysis could be due to the fact that images are usually in halftone format [25] and are difficult to process manually. Even with advanced computational methods, finding features and correlations between different images still remains a challenge. Image data enrichment also has significant value in the same way as text enrichment improves the accessibility and searchability of historical collections [26].

The geolocalization of these historical collections is one of the essential steps to enrich the metadata of historical images. Recent studies show that most geolocalization algorithms are only image-based solutions and work only for well-known landmarks [27]. In this way, many historical images that are less known in the international community but have significant value to a local community cannot be geolocalized. Furthermore, some geolocalization methods are trained on color images and fail to geolocate black and white historical image collections. The main goal of this research was to present a state-of-the-art solution for geolocalizing much more than famous landmarks. Contextual information related to an image is vital for geolocating untagged images. Besides the image features, we also considered the image captions, title, and even the full-text article to extract the location information of a particular historical image. In this way, we could geolocate more historical images than just well-known landmarks.

The next step after geolocalization is an attractive visualization of historical content. For this purpose, we chose the computer-aided rephotography of historical images to show the time travel between two different timestamps of a given location. As the name implies, rephotography is the process of recapturing a photograph from the same point of view [28]. Rephotography is valuable for comparing historical versus recent architectural heritage-

related visual material [29]. Performing rephotography on historical images is notoriously difficult due to the quality of such collections and is usually carried out manually or semiautomatically. Visualizing the temporal evolution in one image can enable humanities scholars to study and analyze changes in a location over time. Some examples of manual rephotography are shown in Figure 1.

Figure 1. Manual rephotography of historic photos and Chicago streets shot during COVID-19 [30,31]. (Credit: Mark Hersch).

This research's results open new directions not only for computer scientists but also for historians and cultural heritage researchers to study spatial and temporal changes in a region or place. The enrichment of metadata facilitates location-based service applications [32]. Geolocalization will be an important feature for a further georeferencing of street-level images [33]. Rephotography will also be of interest to the general public and tourists waiting for a time-based visualization of historical images.

1.1. Dataset

KBR, the Royal Library of Brussels, is the national scientific library of Belgium, responsible for providing access, digitization, and the preservation of all Belgian publications. The library holds over 7 million documents, including newspapers, magazines, books, manuscripts, music, maps, and catalogs [34]. The library not only secures all historical records, but also provides digital access to cultural heritage content to enable historical research. The historical newspaper collection at KBR includes more than 2000 Belgian and 500 international publications. Digitizing and making the historical newspaper collection of KBR accessible has been a long-term project over the last decade [35]. For this study, we used front-page news from the French language socialist newspaper *Le Peuple* from 1938.

1.2. Main Contributions

The main contributions of this research paper are summarized as follows:

- Identification of the location information of a newspaper illustration based on its captions.
- Context-aware geolocalization of building images from KBR's newspaper collections.
- Computational rephotography of historical urban scenes captured at different timestamps.

1.3. General Workflow

This research study aims for the context-aware querying, geolocalization, and rephotography of historical buildings from the KBR digitized newspaper collections. The overview of the general workflow is shown in Figure 2. The first step to geolocate and rephotograph historical images is to digitize newspaper collections. As a result of OCR, an ALTO/XML file is created that contains all the information on a newspaper page (regions where text/illustrations appear, style, alignment information, a link to a paragraph, and description). The next step is to process the newspaper using a document layout analysis approach to segment the articles. Historical images of the buildings are geolocated based on contextual information from the article and illustration captions. A Google image search is used to find similar images based on image descriptions. The irrelevant images are then

filtered out by feature matching algorithms. Finally, images with the same perspective are used to perform the rephotography.

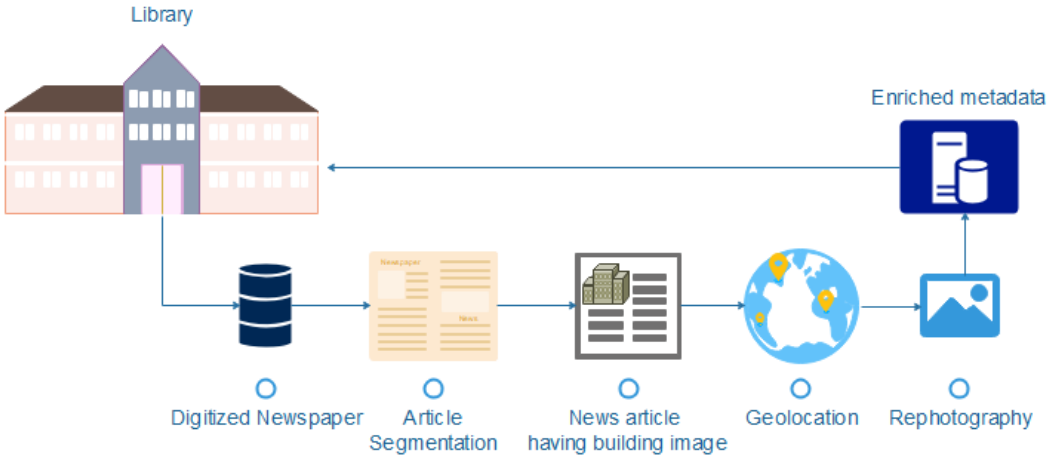


Figure 2. An overview of general workflow.

1.4. Limitations

The quality of OCR (text recognition) must be high enough to provide location information within the proposed workflow. Moreover, if historical photographs contain excessive amounts of noise, geolocalization and rephotography algorithms will not be able to detect appropriate matches.

1.5. Organization of Research Paper

This research paper consists of six sections. Section 2 discusses related research studies. Section 3 presents a methodology for the context-aware querying and geolocalization of historical images. Section 4 discusses feature matching, projective transformations, and the rephotography of historical buildings. Section 5 shows the analysis of the results obtained with our proposed pipeline. Finally, Section 6 draws conclusions.

2. Literature Review

This section summarizes related work in the areas of article-level information retrieval, image feature extraction, matching of features, geolocalization, and visualization of historical image collections.

2.1. Digitization and Extraction

Digital access to cultural heritage materials can be provided through a web interface. Initially, web interfaces did not offer article-level searches, which made it difficult to access relevant data [19]. With an article-level search, one can query a collection based on a particular context, for example, searching for articles in collections related to a specific topic, place, event, or object. The primary purpose of article segmentation is to extract more focused information compared to what is possible with a page-level search. Article segmentation is challenging because of the different layouts, content types, and languages. Many different techniques have been proposed to accomplish this task, such as segmentation based on reading order [36,37], layout information [19,38], segmenting close regions [39], redefining the text blocks [40], and a semantic linking of the textual information [41].

Machine learning approaches have also been used to extract information from historical documents [42–45]. A number of these approaches rely on image-based solutions to classify and categorize them. Several methods are used for the context-aware querying of historical collections, such as clustering [46,47], detecting patterns/similarity [48], and linking articles to open data sources such as Wikipedia and DBpedia [49]. Topic modeling [50] and the clustering of news items on a single topic [51,52], such as politics, sports,

and education, can be used to enrich textual data in newspaper collections. Furthermore, it is now possible to segment historical documents using some recent state-of-the-art machine learning approaches such as PubLayNet [53], DocBank [54], and DocBed [55]. Alternatively, layout analysis models can also be trained on other datasets [56,57]. Compared to standard OCR results, these models achieve good layout segmentation, but they lack the article-level information that is required to improve searchability in historical collections. Just as text data can be classified according to its characteristics, or content, illustrations can also be classified according to their context [58], location, and features such as color or shape [59], enabling the evaluation of visual content.

2.2. Feature Extraction

In computer vision, images are often described as vectors or descriptors that represent the task-specific, distinguishable characteristics of an image. These vectors can be extracted by a variety of different algorithms. In the past, features were extracted manually based on the task. Later, computer vision approaches enabled the extraction of general features such as SIFT [60], SURF [61], or ORB [62]. Nowadays, as deep learning is popular in computer vision domain, convolutional neural networks (CNN) are trained for the identification and feature extraction of specific objects, such as buildings [63,64]. These architectures are typically designed for image classification but can also be adapted for feature extraction. They often have a high-dimensional vector as output which can be reduced to the desired dimension using dimensionality reduction techniques such as a principal component analysis (PCA) [65]. Some models are trained on images of landmarks and buildings to detect distinguishable features between buildings. A currently popular CNN model for feature extraction was published by Radenović et al. [64] under the name Cirtorch. It is a fine-tuned ResNet101 [66] designed for feature extraction from images on which buildings and landmarks are mapped. This model extracts one vector per image, a global descriptor. Other approaches extract the most important keypoints of an image and describe these points in separate vectors, the local descriptors.

2.3. Feature Matching for Geolocalization

A well-known algorithm for finding the most similar vectors is called k-nearest neighbors (k-NN) [67]. The distances between the query vector and the data are calculated to determine the nearest data vectors. Different distance metrics lead to different results. The most popular metric is the Euclidean distance [68]. The distance between two vectors, x and y with dimensionality j can be calculated as in Equation (1) for the Euclidean distance or Equation (2) for the Manhattan distance. When using these metrics, it is important to normalize the vectors when all features in the vectors are equally important.

$$d_{x,y} = \sqrt{\sum_{i=1}^{j}(x_i - y_i)^2} \tag{1}$$

$$d_{x,y} = \sum_{i=1}^{j}|x_i - y_i| \tag{2}$$

A set number of neighbors or a distance threshold determines whether a certain number of matches are found or not. k-NN works with both global and local descriptors. To use the information of local descriptors, a geometric similarity can be determined. Some models, such as DELF [63], extract local features, which means that vectors are extracted per feature in combination with the location of the feature. First, k-NN is used to determine the most similar keypoints. These keypoints can be used to determine the best geometric transformation between the two sets. The transformation is performed by fitting a transformation matrix using RANSAC (random sample consensus) [69]. The number of keypoints that follow the found transformation, called inliers, determines the similarity between images.

2.4. Rephotography

Rephotography is of great value for research in historical studies because it tracks changes over time. In the past, historical scenes were rephotographed manually [70]. Manual rephotography is complex and requires special photographic skills and time to achieve good results. A good rephotograph depends on the precise location, angle, and perspective from which the historical image was taken. Therefore, recapturing a historical scene from the same perspective has been challenging. Computer vision research enables many assisting applications [28] that help the photographer to capture the image from the same viewpoint. Besides these applications, many approaches have been proposed to achieve rephotography based on image collections [71]. The advances in computer technology allow researchers to address this problem using deep learning [29]. The approach of extracting the facade and performing rephotography works well for images of houses where the facade is prominent, but detecting the facade of complex building structures makes this task even more difficult. By using computer vision techniques such as SIFT, ORF, SURF, Cirtorch, or DELF we can solve the problem to some extent, but improvements are still needed to learn and match features from low-quality historical images.

3. Geolocalization of Historical Image Collections

The pipeline for the geolocalization of historical images is explained step by step below.

3.1. Article-Level Segmentation

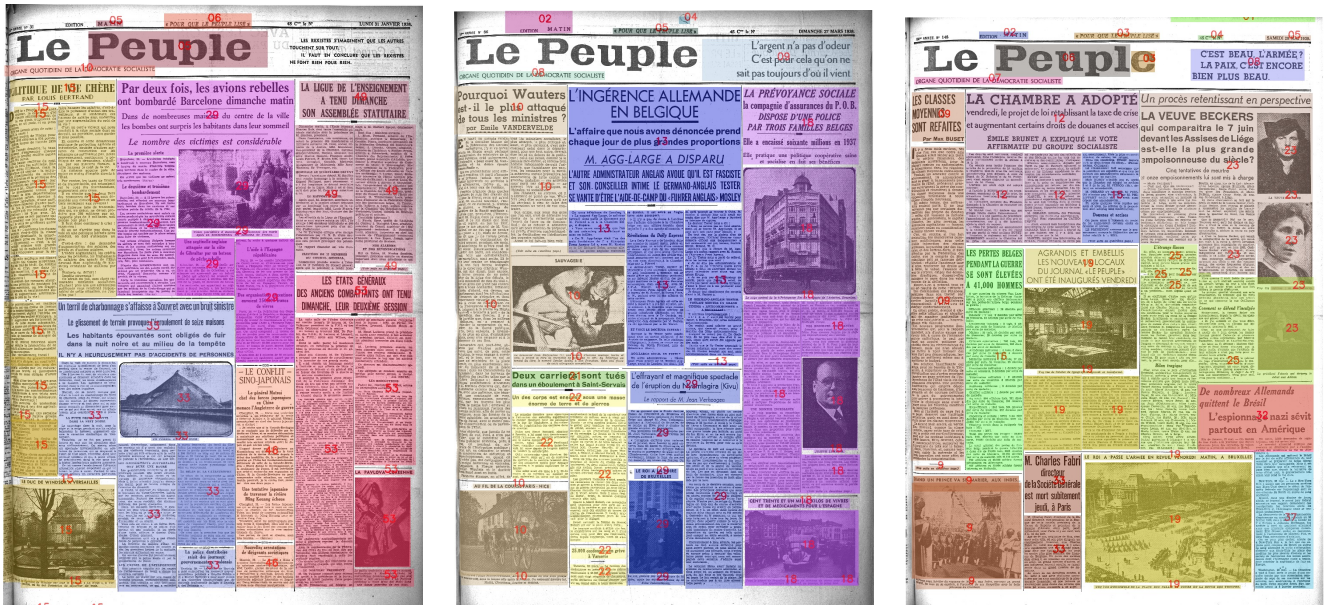
The first step is to gain access to the newspaper collections at the article level. Several algorithms have been proposed to extract the article-level information from historical newspapers. Broadly speaking, there are two ways to extract articles from newspapers: layout-based segmentation and semantic-based segmentation.

The first method is article segmentation based on layout information. In this method, the information about the positioning of the extracted text blocks, the boundary information for the region and the pixel information are used to connect different parts of the article. It is a kind of blind method in which the text within the text blocks is not considered when defining an article. A second method is the semantic linking of information using natural language processing algorithms to create a link between two articles. Several methods are available for extracting articles based on semantics, including fuzzy matching, text classification, and topic model/detection.

Even if all newspapers are printed hierarchically, the layout of a newspaper often changes over time. Therefore, developing a generic algorithm to extract articles can be quite challenging. We developed layout-based methods to segment articles from KBR's digitized newspaper collections. Figure 3 shows some results of our article segmentation using the *Le Peuple* dataset from 1938.

3.2. Context-Aware Article Extraction

In the next step, we extracted only the articles that contained images of a building. We could achieve this using an image classification algorithm. Using historical images, we trained a deep learning algorithm to extract building objects from the historical images. Figure 4 shows how the article text associated with a detected building image is extracted and passed on to the next step for caption detection. To evaluate the proposed approach, we measured the F1 score. We found that the F1 score of the newspaper article segmentation was 0.76.



(**a**) 31st Jan 1938 (**b**) 27th March 1938 (**c**) 28th May 1938

Figure 3. Article segmentation results on Le Peuple 1938 dataset.

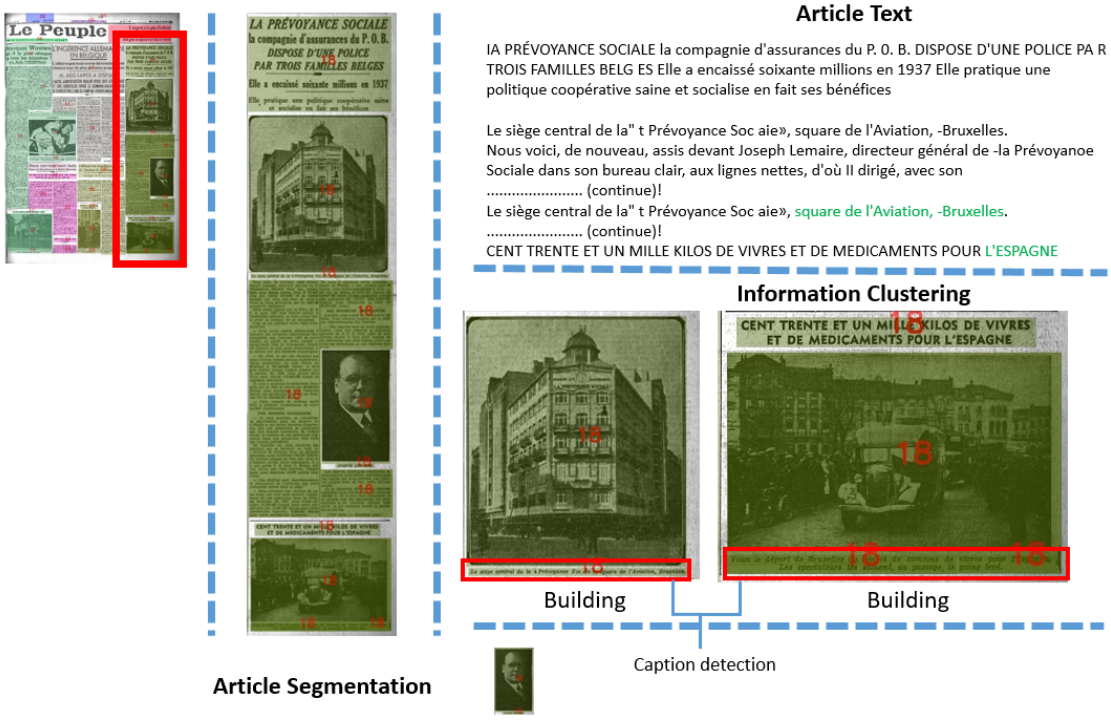


Figure 4. An example of article with a building illustration

3.3. Caption Detection

Captions are the textual information that typically appears below an illustration and describes the illustration. The location information in the caption can be used not only to improve the searchability of content, but also to link photographs to articles and Wikidata entries. However, it can also be used to geolocate historical photographs. Some features help identify captions. These features include text blocks with the same width as the image, placed directly below the image, and styled differently than normal text. There are red blocks underneath the illustrations in Figure 4 that indicate detected captions.

3.4. Location Information Extraction

Geographical information is sometimes provided in the captions, which aids in geolocating the historical images. The captions of an illustration or article text usually contain information about the event or location. In this step, natural language techniques (NER methods [72]) were applied to extract the location information from the article caption or text. The location information could be a place name, a city name, or a country name. An example of location information extraction from the text is shown below.

Caption tag: Le siège central de la" t Prévoyance Soc aie», square de l'Aviation, -Bruxelles.

Location information: 'Prévoyance Sociale [LOC], square de l'Aviation, -Bruxelles [LOC].

When extracting location information, we encountered difficulties with poor OCR quality and multiple location detection. The location cannot be determined if the OCR quality is not high enough to provide enough information about the location. When we had more than one location in the caption or article text, we matched historical image features with Google images to determine the exact location.

3.5. Geolocalization

The final step was geolocalization, which is a two-step procedure. In the first step, we searched for relevant images using a Google image search based on the location information, extracted from the caption. For each image, we extracted the top 10 results. The Google search gave all correct results when we searched for well-known landmarks, e.g., *"the Castle of Prague"*, but in the case of a less common landmark, the Google search results always returned some irrelevant images, e.g., the search result of *"Femmes Prévoyantes du Centre, Nieuport Bains"* shows only two correct results for the mentioned building. This study aimed to be able to geolocate less-common images of buildings in addition to the well-known landmarks. Therefore, in the second step, we needed to filter the search results by keywords to find out which images were similar to the image shown in the newspaper. Our pipeline for finding the best matches is shown in Figure 5. We computed the features of all images and compared the similarity score. Using a threshold, we filtered out the irrelevant results.

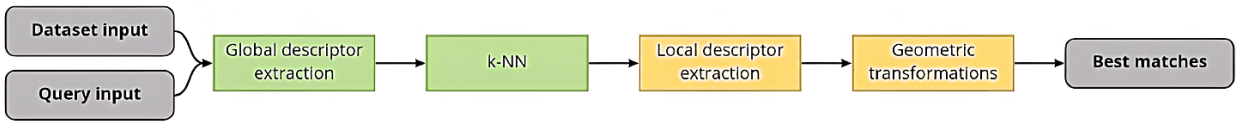


Figure 5. Pipeline to find the best matches

CNNs designed for extracting features from landmarks or buildings are usually trained on contemporary color images. These do not perform well on monochrome (black-and-white) and halftone images. The proposed pipeline for building matching hierarchically combined global and local descriptors extracted using the pretrained Cirtorch and DELF models. In this way, the results were improved for halftone newspaper images compared to matching the descriptors separately.

The feature matching algorithm first extracted the global descriptors for all images in the database using Cirtorch. These global descriptors were used to filter out the irrelevant images using k-NN. Figure 6 shows the filtering of irrelevant results. Most of the downloaded images provided good results, hence only the three (3) worst images were excluded from the next step of the pipeline, local descriptor matching. Local descriptors were extracted from the seven (7) remaining images using DELF. From this model, not only the features for the keypoints were retrieved, but also the location of each one. This allowed us to search for similar images based on geometric transformations of the keypoints using RANSAC. Images with a confidence score greater than 0.5 were considered good matches.

Figure 6. Search results for similar images using caption tags and filtering irrelevant images based on feature matching

4. Rephotography of Historical Image Collections

The term "rephotography" refers to the practice of taking the same subject at two different timestamps. Computational rephotography can be challenging due to the poor quality of the original historical photographs. Furthermore, the angle and location from which the images were taken, as well as the perspective and the changes in the building itself over time can also affect the outcomes of the rephotography. Pictures from historical newspapers and similar images found through a Google search helped find suitable images for the rephotography task. SIFT/ORB was the computer vision-based method used for detecting, describing, and matching local features in historical images.

We used images taken at the same location at two different timestamps for the rephotography. One image was the source image (historical image), as shown in Figure 7a, and the other was the target image (relatively new/another view of the same location) as shown in Figure 7b. Using feature matching algorithms, we could identify the common keypoints to both images as shown in Figure 8a. Using these keypoints, we could find a suitable transformation matrix from the source image to the target image. As shown in Figure 8b, a perspective transformation was applied to the source image to get the same viewpoint for both images. The next step was to determine an appropriate crop in the source image to be pasted on the target image (see Figure 8c). In the final step, some blending and blurring effects were applied in the region where both images were stitched to show the smooth transition of a scene from the historical image to the current view, as shown in Figure 8d. Some more results can be seen in Figures 9 and 10.

(**a**) Historical Image (**b**) Recent view of same location

Figure 7. Search results for similar images using caption tags.

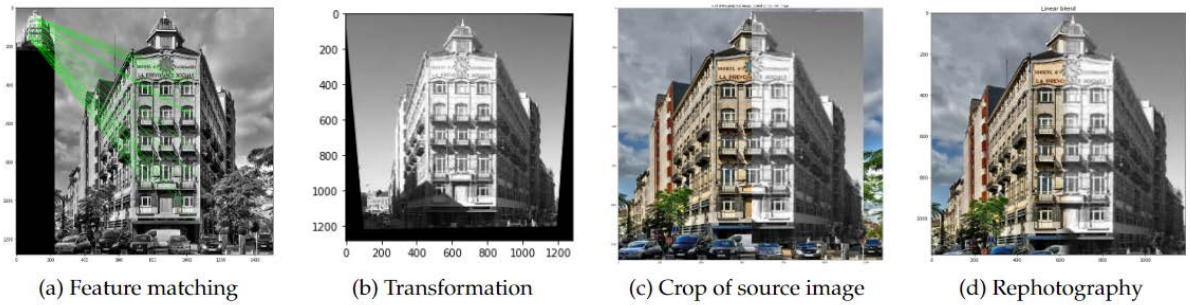


Figure 8. Rephotography of historical images.

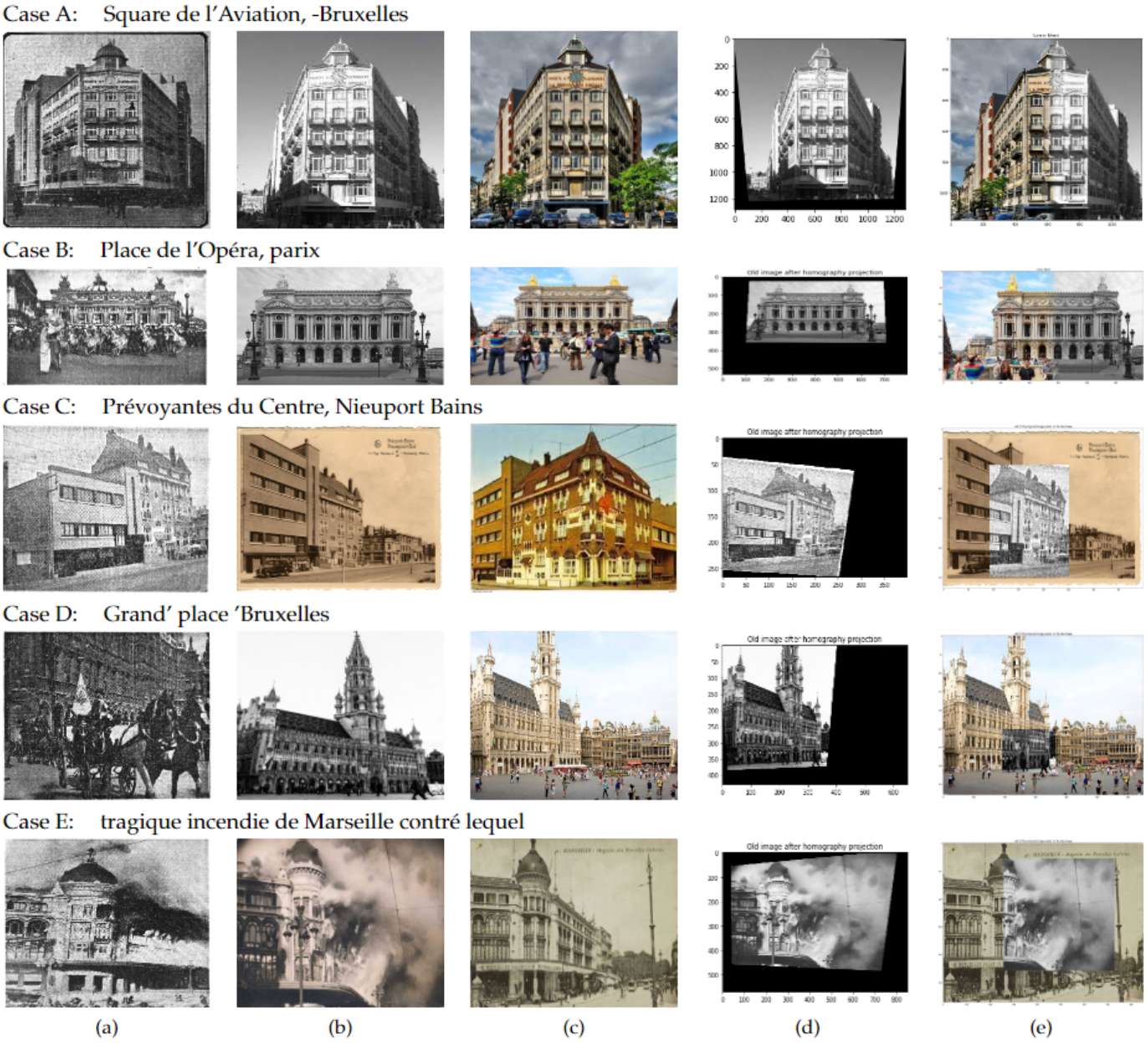
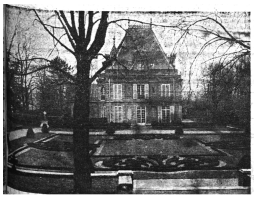


Figure 9. Image extraction and rephotography using location information: (**a**) historical image; (**b**) extracted image 1; (**c**) extracted image 2; (**d**) transformation; (**e**) rephotography.

Case F: ’ville mexicaine de taxco catholique

Case G: l’’Institut d’’histoire sociale, Prévoyance Sociale, boulevard de le Régent

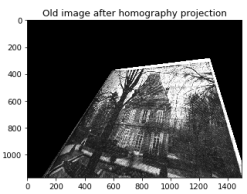
Case H: Duc de Windsor, château de maye, Versailles



(a)

(b)

(c)



(d)

(e)

Figure 10. Some failed cases : (**a**) historical image; (**b**) extracted image 1; (**c**) extracted image 2; (**d**) transformation; (**e**) rephotography.

5. Results and Discussion

The aim of this study was to present a novel approach for exploring cultural heritage datasets. This exploratory study proposed an entire pipeline for geolocalization and rephotography of historical images. The dataset consisted of 40 articles with buildings based on front-page news from the 1938 French language socialist newspaper *Le Peuple*. First, we processed these 40 images by our location extraction algorithm. There were eight instances where the OCR results did not detect the location information. The location information of the remaining 32 articles was correctly detected. In the next step, we applied a Google query based on location information and filtered irrelevant results based on feature matching. When searching the Internet for similar perspectives, 6 out of the remaining 32 instances failed to find relevant perspectives. The final step in our pipeline was the rephotography of similar-perspective images. Of the remaining 26 instances, 5 did not find the appropriate transformation in this step. The remaining 21 were successfully rephotographed. In Table 1, we demonstrate the accuracy of the proposed pipeline on the basis of 40 articles with building information. Our results indicated that we achieved an average accuracy of 80% for each step in the proposed pipeline. The overall accuracy of the entire pipeline was 52.5%. This means that every second historical image passed successfully through the pipeline and was rephotographed. Given the difficulty of this research area, this is considered a substantial contribution and a notable success. Since this is the first computational rephotography pipeline using traditional computer vision methods, this research opens a new direction for research with historical images. The entire pipeline was developed as an exploratory study, so each section needs further revision to improve accuracy, thereby increasing the overall accuracy of the pipeline. A detailed study of each section with a larger dataset will be conducted in the future to improve the accuracy of the pipeline.

Furthermore, this section discusses the cases we encountered during our research using the proposed workflow. Figure 9 shows the results of our complete workflow to geolocalize and visualize the historical collections of KBR. In this section, we discuss eight different cases. In case A, B, C, D, and E, it can be clearly seen that the extracted images

matched very well with the reference historical images. The feature matching pipeline could detect good features, resulting in astonishing transformation and rephotography.

In case A *"Square de l'Aviation, -Bruxelles"* and case B *"Place de l'Opéra, parix"* the rephotography was performed on the extracted image sets with different timestamps. It is necessary to apply an image smoothing algorithm in order to rephotograph historical photographs. Case C *"Prévoyantes du Centre, Nieuport Bains"* shows an exceptional match between the historical image and the picture of the same location from different timestamps. Case D *"Grand' place 'Bruxelles"* shows the rephotography of a part of the image. In this case, only the features at the facade plane matched correctly, so the rest of the features were ignored during the rephotography step. Case E *"tragique incendie de Marseille contré lequel"* refers to scenarios where the location we detected was too general. The detected location was *"Marseille"*, which was too general information about a location to search for the building. Therefore, we added additional information about the event/incident that happened at *Marseille*. This contextual information helped extract the corresponding images. Finally, the rephotography of this scene shows the picture before and during the fire incident in a single picture.

Table 1. Accuracy of proposed pipeline based on 40 articles with buildings.

	Each Subsection			Full Pipeline
	Location Information	**Geolocalization**	**Rephotography**	
Correctly processed	32	26	21	
Accuracy per section	80%	81%	81%	
Accuracy full pipeline				52.5%

Some false cases (case F, G, and H) are also shown in Figure 10 . Case F *"'wsi. [mexicaine de taxco catholique]"* failed to detect similar buildings because of the poor quality of the OCR results. In this case, the caption tags detected using OCR were *"•3?HH - **?*. /- wsi.: .. . .CTagfigaj *•****»: "*, which made it impossible to extract the correct location *"ville mexicaine de taxco catholique"* from the caption. Hence, if the OCR is of poor quality, the location information extraction algorithm is not be able to detect the correct location. This makes it impossible to proceed with the geolocalization and rephotography steps. In addition, the proposed method may fail if we perform a keyword search and cannot locate a matching building image. Case G *"l"Institut d"histoire sociale, Prévoyance Sociale, boulevard de le Régent"* shows that the text detection was good, but no matching building was found to proceed with the geolocalization and rephotography steps.

Another challenge that might arise with this workflow is the fact that the feature matching algorithm may not always be able to identify suitable matches to perform good rephotography. Case H *"Duc de Windsor, château de maye, Versailles"* shows the scenario where we found a similar building but failed to perform rephotography of the scene. The size of the picture, picture quality, the partial occlusion due to the environment, significant changes in the location, and the completely different perspective from the reference image are some of the factors that can affect the rephotography results.

6. Conclusions

In this research study, we proposed a methodology for the geolocalization and visualization of historical images from KBR's newspaper collections. Nowadays, computer vision and machine learning approaches are mature enough to explore massive library collections more efficiently and in less time. Article-level information enables users to query and explore the KBR historical collection on a particular subject. Natural language processing was used to retrieve context from an article's heading, text, or illustration caption. The context could be location information, a person's name, or an event description. Any hint to the location information was valuable for geolocating the untagged images. A keyword-based search for location information found all the associated images. Feature

matching techniques further filtered out the irrelevant results. In the final step, the best matched geolocalized images were used to perform the rephotography.

Contextual information is the detail/description of an image and serves as additional information that leads to the metadata enrichment of historical collections. The enriched metadata further enhance digital access and searchability within the historical collection. Geolocalization and rephotography are both tasks that require excellent OCR accuracy, which is one of the common challenges associated with digitizing documents from the past. The computer vision approach to computer-aided content analysis makes it easier to propose a solution that allows users to query, geolocate, and rephotograph portions of historical collections.

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

Funding: This research has been funded by the DATA-KBR-BE project (2020-2024) financed by the Belgian Science Policy Office (Belspo) as part of the Belgian Research Action through Interdisciplinary Networks, BRAIN 2.0 program which is coordinated by KBR.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank the KBR for enabling access to the historical newspaper data for this research.

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

References

1. Batista, G.E.; Monard, M.C. A study of K-nearest neighbour as an imputation method. *His* **2002**, *87*, 48.
2. Hughes, L.M. *Digitizing Collections: Strategic Issues for the Information Manager*; Facet Publishing: London, UK, 2004; Volume 2.
3. Nielsen, E.K. Digitisation of library material in Europe: Problems, obstacles and perspectives anno 2007. *Liber Q. J. Assoc. Eur. Res. Libr.* **2008**, *18*, 20–27. [CrossRef]
4. Ehrmann, M.; Bunout, E.; Duering, M. *Survey of Digitized Newspaper Interfaces*; InfoScience: Lausanne, Switzerland, 2019.
5. Bingham, A. The digitization of newspaper archives: Opportunities and challenges for historians. *Twent. Century Br. Hist.* **2010**, *21*, 225–231. [CrossRef]
6. Traub, M.C.; Ossenbruggen, J.v.; Hardman, L. Impact analysis of OCR quality on research tasks in digital archives. In Proceedings of the International Conference on Theory and Practice of Digital Libraries, , Poznań, Poland, 14–19 September 2015; pp. 252–263.
7. Ali, D.; Verstockt, S. Challenges in Extraction and Classification of News Articles from Historical Newspapers. In Proceedings of the What is Past is Prologue: The NewsEye International Conference 2021, Virtual, 16–17 March 2021; pp. 8–9.
8. James-Gilboe, L. The challenge of digitization: Libraries are finding that newspaper projects are not for the faint of heart. *Ser. Libr.* **2005**, *49*, 155–163. [CrossRef]
9. Hill, M.J.; Hengchen, S. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digit. Scholarsh. Humanit.* **2019**, *34*, 825–843. [CrossRef]
10. Jarlbrink, J.; Snickars, P. Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive. *J. Doc.* **2017**, *73*, 1228–1243. [CrossRef]
11. Fugini, M.; Finocchi, J. Data and Process Quality Evaluation in a Textual Big Data Archiving System. *ACM J. Comput. Cult. Herit.* **2022**, *15*, 1–19. [CrossRef]
12. Lu, T.; Ilic, D.; Dooms, A. Noise characterization for historical documents with physical distortions. In Proceedings of the Optics, Photonics and Digital Technologies for Imaging Applications VI. International Society for Optics and Photonics, Online, 6–10 April 2020; Volume 11353, p. 113530F.
13. BelgicaPress Platform. Available online: https://www.belgicapress.be/?lang=EN (accessed on 26 May 2022).
14. Michalakis, K.; Caridakis, G. Context Awareness in Cultural Heritage Applications: A Survey. *ACM J. Comput. Cult. Herit.* **2022**, *15*, 1–31. [CrossRef]
15. Dahroug, A.; Vlachidis, A.; Liapis, A.; Bikakis, A.; Lopez-Nores, M.; Sacco, O.; Pazos-Arias, J.J. Using dates as contextual information for personalised cultural heritage experiences. *J. Inf. Sci.* **2021**, *47*, 82–100. [CrossRef]

16. Chaudhury, K.; Jain, A.; Thirthala, S.; Sahasranaman, V.; Saxena, S.; Mahalingam, S. Google newspaper search–image processing and analysis pipeline. In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 621–625.
17. Pekárek, A.; Willems, M. The europeana newspapers—A gateway to european newspapers online. In Proceedings of the Euro-Mediterranean Conference, Limassol, Cyprus, 29 October–3 November 2012; pp. 654–659.
18. Lee, B.C.G.; Mears, J.; Jakeway, E.; Ferriter, M.; Adams, C.; Yarasavage, N.; Thomas, D.; Zwaard, K.; Weld, D.S. The newspaper navigator dataset: Extracting and analyzing visual content from 16 million historic newspaper pages in chronicling America. *arXiv* **2020**, arXiv:2005.01583.
19. Ali, D.; Millivellie, K.; van den Broeck, A.; Verstockt, S. NewspAIper: AI-Based Metadata Enrichment of Historical Newspaper Collections. 2022. Available online: https://zenodo.org/record/6593028/export/dcat (accessed on 23 October 2022).
20. Kemman, M.; Claeyssens, S. User demand for supporting advanced analysis of historical text collections. In Proceedings of the DH Benelux 2022—ReMIX: Creation and alteration in DH (hybrid), Belval Campus, Esch-sur-Alzette, Luxembourg, Online, 1–3 June 2022. [CrossRef]
21. Kestemont, M.; Karsdorp, F.; Düring, M. Mining the twentieth century's history from the time magazine corpus. In Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), Gothenburg, Sweden, 26 April 2014; pp. 62–70.
22. Bikakis, A.; Hyvönen, E.; Jean, S.; Markhoff, B.; Mosca, A. Special issue on Semantic Web for Cultural Heritage. *Semant. Web* **2021**, *12*, 163–167. [CrossRef]
23. Baqués, X.G.; Mosca, A.; Rondelli, B.; Fort, G.R. Roman Open Data: A semantic based Data Visualization & Exploratory Interface.
24. Roffo, G.; Giorgetta, C.; Ferrario, R.; Riviera, W.; Cristani, M. Statistical analysis of personality and identity in chats using a keylogging platform. In Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, 12–16 November 2014; pp. 224–231.
25. Gao, Q.; Shu, X.; Wu, X. Deep restoration of vintage photographs from scanned halftone prints. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 4120–4129.
26. Smits, T. Illustrations to Photographs: Using computer vision to analyse news pictures in Dutch newspapers, 1860–1940. In Proceedings of the DH, Montréal, QC, Canada, 8–11 August 2017.
27. Tahmasebzadeh, G.; Kacupaj, E.; Müller-Budack, E.; Hakimov, S.; Lehmann, J.; Ewerth, R. GeoWINE: Geolocation based Wiki, Image, News and Event Retrieval. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 11–15 July 2021; pp. 2565–2569.
28. Bae, S.; Agarwala, A.; Durand, F. Computational rephotography. *ACM Trans. Graph.* **2010**, *29*, 24. [CrossRef]
29. Ali, D.; Verstockt, S.; Van de Weghe, N. Single Image Façade Segmentation and Computational Rephotography of House Images Using Deep Learning. *J. Comput. Cult. Herit.* **2021**, *14*, 1–17. [CrossRef]
30. Idowu, A. How 'Rephotography' is Capturing Chicago in the Age of COVID-19. 2020. Available online: https://news.wttw.com/2020/04/21/how-rephotography-capturing-chicago-age-covid-19 (accessed on 6 June 2022).
31. Hersch, M. Time After Time, Rephotography in the Midst of a Pandemic. 2020. Available online: https://www.markhersch.com/rephotography-in-the-midst-of-a-pandemic (accessed on 6 June 2022).
32. Huang, H.; Gartner, G.; Krisp, J.M.; Raubal, M.; Van de Weghe, N. Location based services: Ongoing evolution and research agenda. *J. Locat. Based Serv.* **2018**, *12*, 63–93. [CrossRef]
33. Harrach, M.; Devaux, A.; Brédif, M. Interactive image geolocalization in an immersive web application. *Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2019**, *42*, 377–380. [CrossRef]
34. KBR Collections. Available online: https://www.kbr.be/en/collections/ (accessed on 26 May 2022).
35. Chambers, S.; Lemmers, F.; Pham, T.A.; Birkholz, J.M.; Ducatteeuw, V.; Jacquet, A.; Dillen, W.; Ali, D.; Milleville, K.; Verstockt, S. Collections as Data: Interdisciplinary experiments with KBR's digitised historical newspapers: A Belgian case study. In Proceedings of the 7th DH Benelux: The Humanities in a Digital World (DH Benelux 2021), Belgium, The Netherlands, 2–4 June 2021.
36. Gao, L.; Tang, Z.; Lin, X.; Wang, Y. A graph-based method of newspaper article reconstruction. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 1566–1569.
37. Naoum, A.; Nothman, J.; Curran, J. Article segmentation in digitised newspapers with a 2D Markov model. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 1007–1014.
38. Chen, M.; Ding, X.; Liang, J. Analysis, Understanding and Representation of Chinese newspaper with complex layout. In Proceedings of the Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101), Vancouver, BC, Canada, 10–13 September 2000; Volume 2, pp. 590–593.
39. Reynar, J.C. An automatic method of finding topic boundaries. *arXiv* **1994**, arXiv:cmp-lg/9406017.
40. Hadjar, K.; Hitz, O.; Ingold, R. Newspaper page decomposition using a split and merge approach. In Proceedings of the Sixth International Conference on Document Analysis and Recognition, Seattle, WA, USA, 13 September 2001; pp. 1186–1189.
41. Doerr, M.; Markakis, G.; Theodoridou, M.; Tsikritzis, M. DIATHESIS: OCR based semantic annotation of newspapers. In Proceedings of the Third SEEDI International Conference: Digitization of Cultural and Scientific Heritage, Cetinje, Montenegro, 13–15 September 2007.

42. Meier, B.; Stadelmann, T.; Stampfli, J.; Arnold, M.; Cieliebak, M. Fully convolutional neural networks for newspaper article segmentation. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 414–419.
43. Paaß, G.; Konya, I. Machine learning for document structure recognition. In *Modeling, Learning, and Processing of Text Technological Data Structures*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 221–247.
44. Almutairi, A.; Almashan, M. Instance segmentation of newspaper elements using mask R-CNN. In Proceedings of the 2019 18th IEEE International Conference On Machine Learning Furthermore, Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 1371–1375.
45. Bhatt, J.; Hashmi, K.A.; Afzal, M.Z.; Stricker, D. A survey of graphical page object detection with deep neural networks. *Appl. Sci.* **2021**, *11*, 5344. [CrossRef]
46. Aggarwal, C.C.; Zhai, C. A survey of text clustering algorithms. In *Mining Text Data*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 77–128.
47. Aggarwal, C.C.; Reddy, C.K. Data clustering. *Algorithms and Applications*; Chapman&Hall/CRC Data Mining and Knowledge Discovery Series; Chapman&Hall/CRC: London, UK, 2014.
48. Aletras, N.; Stevenson, M.; Clough, P. Computing similarity between items in a digital library of cultural heritage. *J. Comput. Cult. Herit.* **2013**, *5*, 1–19. [CrossRef]
49. Linhares Pontes, E.; Cabrera-Diego, L.A.; Moreno, J.G.; Boros, E.; Hamdi, A.; Sidère, N.; Coustaty, M.; Doucet, A. Entity Linking for Historical Documents: Challenges and Solutions. In Proceedings of the International Conference on Asian Digital Libraries, Kyoto, Japan, 30 November–1 December 2020; pp. 215–231.
50. DiMaggio, P.; Nag, M.; Blei, D. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics* **2013**, *41*, 570–606. [CrossRef]
51. Aker, A.; Kurtic, E.; Balamurali, A.; Paramita, M.; Barker, E.; Hepple, M.; Gaizauskas, R. A graph-based approach to topic clustering for online comments to news. In Proceedings of the European Conference on Information Retrieval, Padua, Italy, 20–23 March 2016; pp. 15–29.
52. Mele, I.; Bahrainian, S.A.; Crestani, F. Event mining and timeliness analysis from heterogeneous news streams. *Inf. Process. Manag.* **2019**, *56*, 969–993. [CrossRef]
53. Zhong, X.; Tang, J.; Jimeno Yepes, A. PubLayNet: Largest dataset ever for document layout analysis. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019.
54. Li, M.; Xu, Y.; Cui, L.; Huang, S.; Wei, F.; Li, Z.; Zhou, M. DocBank: A Benchmark Dataset for Document Layout Analysis. *arXiv* **2020**, arXiv:cs.CL/2006.01038.
55. Zhu, W.; Sokhandan, N.; Yang, G.; Martin, S.; Sathyanarayana, S. DocBed: A multi-stage OCR solution for documents with complex layouts. *Proc. Conf. AAAI Artif. Intell.* **2022**, *36*, 12643–12649. [CrossRef]
56. Clausner, C.; Papadopoulos, C.; Pletschacher, S.; Antonacopoulos, A. The ENP image and ground truth dataset of historical newspapers. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015.
57. Antonacopoulos, A.; Clausner, C.; Papadopoulos, C.; Pletschacher, S. Icdar 2013 competition on historical newspaper layout analysis (hnla). In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1454–1458.
58. Düring, M.; Kalyakin, R.; Bunout, E.; Guido, D. Impresso inspect and Compare. Visual comparison of semantically enriched historical newspaper articles. *Information* **2021**, *12*, 348. [CrossRef]
59. Manovich, L. Data science and digital art history. *Int. J. Digit. Art Hist.* **2015**. [CrossRef]
60. Lowe, G. Sift-the scale invariant feature transform. *Int. J.* **2004,** *2*, 91–110.
61. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
62. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Colorado Springs, CO, USA, 20–25 June 2011; pp. 2564–2571.
63. Noh, H.; Araujo, A.; Sim, J.; Weyand, T.; Han, B. Large-scale image retrieval with attentive deep local features. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3456–3465.
64. Radenović, F.; Tolias, G.; Chum, O. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 3–20.
65. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
66. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029.
67. Cunningham, P.; Delany, S.J. k-Nearest neighbour classifiers-A Tutorial. *ACM Comput. Surv.* **2021**, *54*, 1–25. [CrossRef]
68. Wang, L.; Zhang, Y.; Feng, J. On the Euclidean distance of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1334–1339. [CrossRef]
69. Fischler, M.A.; Bolles, R.C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
70. Mitotes, W. Then and Now: The Time Portals into a World War II. 2010. Available online: http://weramitotes.blogspot.com/2010/08/then-and-now-time-portals-into-world.html (accessed on 26 May 2022).

71. Lee, K.T.; Luo, S.J.; Chen, B.Y. Rephotography using image collections. *Comput. Graph. Forum* **2011**, *30*, 1895–1901. [CrossRef]
72. Mac Kim, S.; Cassidy, S. Finding names in trove: Named entity recognition for Australian historical newspapers. In Proceedings of the Australasian Language Technology Association Workshop 2015, Melbourne, Australia, 8–9 December 2015; pp. 57–65.